BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

RESPONSE

I. Status of the Claims

Claims 1 and 2 have been cancelled entirely without prejudice and without disclaimer. New claims 5-12 have been added to better claim the present invention. As a result, claims 3-12 are therefore pending in the present case.

II. Support for the Claims

New claim 5 has been added to more clearly claim aspects of the invention. Claim 5 finds support throughout the specification and sequence listing as originally filed, with particular support being found in original claim 3 on which it depends and on original SEQ ID NO: 1.

New claim 6 has been added to more clearly claim aspects of the invention. Claim 6 finds support throughout the specification and sequence listing as originally filed, with particular support being found in original claim 4 on which it depends and on original SEQ ID NOS: 3.

New claim 7 has been added to more clearly claim aspects of the invention. Claim 7 finds support throughout the specification and sequence listing as originally filed, with particular support being found at least at page 14, lines 24-31 and in original SEQ ID NOS:2 and 4.

New claims 8 and 9 have been added to more clearly claim aspects of the invention. Claims 8 and 9 find support throughout the specification and sequence listing as originally filed, with particular support being found in claim 7 on which they depend and in original SEQ ID NOS:1 and 3, respectively.

New claim 10 has been added to more clearly claim aspects of the invention. Claim 10 finds support throughout the specification and sequence listing as originally filed, with particular support being found at least at page 14, line 31 through page 15, line 4 and in original SEQ ID NOS:2 and 4.

New claim 11 has been added to more clearly claim aspects of the invention. Claim 11 finds support throughout the specification and sequence listing as originally filed, with particular support being found in claim 7 on which it depend and in original SEQ ID NO:1.

New Claim 12 has been added to more clearly claim aspects of the invention. Claim 12 finds support throughout the specification and sequence listing as originally filed, with particular support being found in original claim 7 on which it depends and on original SEQ ID NOS:3.

As new claim 5-12 are fully supported by the specification, sequence listing and claims as originally filed, they do not constitute new matter. Entry is therefore respectfully requested.

III. Rejection of Claims Under 35 U.S.C. § 101

The Action rejects claims under 35 U.S.C. § 101, allegedly because the claimed invention lacks support by either a specific and substantial asserted utility or a well established utility. Applicants respectfully traverse.

The Action discounts many of the numerous utilities described in the specification for the sequences of the present invention based on the position that while credible, these utilities are not specific or substantial. While Applicants in no way agree with the Examiner's arguments, Applicants have chosen to expand on only a few of the utilities as only one is required.

Applicants respectfully submit that the legal test for utility involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be credible or believable. According to the Examination Guidelines for the Utility Requirement, if the applicant has asserted that the claimed invention is useful for any particular purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, the Examiner should not impose a rejection based on lack of utility (66 Federal Register 1098, January 5, 2001).

In *In re Brana,* (34 USPQ2d 1436 (Fed. Cir. 1995), "*Brana*"), the Federal Circuit admonished the P.T.O. for confusing "the requirements under the law for obtaining a patent with the requirements for obtaining government approval to market a particular drug for human consumption". *Brana* at 1442. The Federal Circuit went on to state:

> At issue in this case is an important question of the legal constraints on patent office examination practice and policy. The question is, with regard to pharmaceutical inventions, what must the applicant provide regarding the practical utility or usefulness of the invention for which patent protection is sought. <u>This is not a new issue; it is one which we would have thought had been settled by case law years ago</u>.

Brana at 1439, emphasis added. The choice of the phrase "utility or usefulness" in the foregoing quotation is highly pertinent. The Federal Circuit is evidently using "utility" to refer to rejections under 35 U.S.C. § 101, and is using "usefulness" to refer to rejections under 35 U.S.C. § 112, first

paragraph. This is made evident in the continuing text in *Brana*, which explains the correlation between 35 U.S.C. §§ 101 and 112, first paragraph. The Federal Circuit concluded:

> FDA approval, however, is not a prerequisite for finding a compound useful within the meaning of the patent laws. Usefulness in patent law, and in particular in the context of pharmaceutical inventions, necessarily includes the expectation of further research and development. The stage at which an invention in this field becomes useful is well before it is ready to be administered to humans. Were we to require Phase II testing in order to prove utility, the associated costs would prevent many companies from obtaining patent protection on promising new inventions, thereby eliminating an incentive to pursue, through research and development, potential cures in many crucial areas such as the treatment of cancer.

Brana at 1442-1443, citations omitted. In assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not "experimentation". *In re Angstadt and Griffin*, 190 USPQ 214 (C.C.P.A. 1976). The need for some experimentation does not render the claimed invention unpatentable. Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin, supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 18 USPQ2d 1016 (Fed. Cir. 1991). As a matter of law, it is well settled that a patent need not disclose what is well known in the art. *In re Wands*, 8 USPQ 2d 1400 (Fed. Cir. 1988).

Even under the newly installed utility guidelines, Applicants note that MPEP 2107 (II)(B)(1) states:

> (1) If the applicant has asserted that the claimed invention is useful for any particular practical purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, do not impose a rejection based on lack of utility. (MPEP 2107 (II)(B)(1))

Applicants have asserted that the sequences of the present invention encode a novel human membrane protein containing EGF domains. The Examiner agrees that this assertion is credible, "The assertion that SEQ ID NO:1 encodes a novel EGF is credible" (Action at page 3, lines 11-12). Presumptively this is based on the 100% homology that exists between SEQ ID NO:2 and MEGF10 described in Nagase, *et al.*(2001, cited in the Action). Calcium-binding EGF-like domains are present in membrane-bound and extracellular animal proteins. Many of these proteins require calcium for their

biological function and calcium-binding sites have been found to be located at the N-terminus of particular EGF-like domains; calcium-binding can be crucial for numerous protein-protein interactions. Analysis of the protein encoded by the sequences of the present invention (MEGF10) indicates that it contains Ca2+-binding EGF-like domains and, like many EGF proteins, functions in signal transduction via protein phosphorylation and clathrin-mediated endocytosis leading to gene activation following growth factor binding. Clearly this information allows one of skill in the art to use the sequences of the present invention.

The Action's statements (page 3, lines 6-7) that "The specification discloses no data for any activity of SEQ ID NO:1. There are no working examples" indicating a need for such information are misplaced. It has long been established that "there is no statutory requirement for the disclosure of a specific example". *In re Gay*, 135 USPQ 311 (C.C.P.A. 1962). The Action goes on to state that the invention lacks utility because the disclosure provides no guidance as to where the important structural elements of the protein encoded by the sequences of the present invention that are essential to its function are located (Action at page 4). This concern is also misplaced, as it is well established that "an inventor is not required to understand the theory of how his invention works". *Micro Motion, Inc. v. Exac Corp.*, 16 USPQ2d 1001, 1013 (Cal. 1990).

While having accepted as credible Applicant's assertion that SEQ ID NO:1 encodes a novel EGF protein based on sequence identity and the presence of EGF domains, the Action then contradicts itself by going on to assert extensively that one can deduce accurately a molecules function based on structure (pages 4-5). The Examiner's position appears to be that sequence homology and the relationship between structure and function is not generally accepted by those of skill in the art. In support of this position the Action goes on to describe again several contrarian articles that the Action alleges indicate that "the art recognizes that function cannot be predicted from structure alone"

The Examiner cites Bork (Genome Research *10*:398-400, 2000) as supporting the proposition that prediction of protein function from homology information is somewhat unpredictable. It is of interest that in his "analysis" Bork often uses citations to many of his own previous publications, an interesting approach. ' My position is supported by my previous disclosures of my position.' If Bork's position is supported by others of skill in the art, one would expect that he would reference them rather than himself to provide support for his statements. Given that the standard with regard to obtaining U.S. patents is those of skill in the art, this observation casts doubt on the broad applicability of Bork's

position. I t should also be noted that in Table 1, on page 399, in which selected examples of prediction accuracy are presented, that the reported accuracy of the methods which Appellants have employed are, in fact, very high. While nowhere in Bork is there a comparison of the prediction accuracy based on the percentage homology between two proteins or two classes of proteins, " Homology (several methods)" is assigned an accuracy rate of 98% and "Functional features by homology" is assigned an accuracy rate of 90%. Given that these figures were obtained based on what is at least a 4 year old analysis, these high levels of accuracy would appear to support rather than refute Appellants assertions in the present case. Additionally Bork even states (on page 400, second column, line 17) that " However, there is still no doubt that sequence analysis is extremely powerful". In summary, it is clear that it is not Bork's intention to refute the value of sequence analysis but rather he is indicating that there is room for improvement.

The Action also cites an article by Skolnick and Fetrow ("Skolnick"; 2000, TIBTECH 18:34-39) for the proposition that "(k)nowing the protein structure by itself is insufficient to annotate a number of functional classes and is also insufficient for annotating the specific details of protein function" (Skolnick at page 36, emphasis added). However, Skolnick concerns predicting protein function not by overall amino acid homology to other family members, but instead concerns prediction of function based on the presence of certain functional "motifs" present within a given protein sequence. Thus, Skolnick does not apply to the current situation, where overall protein homology is used, as described by Ji, to assign function to a particular sequence. However, even in the event that Skolnick is applicable, Skolnick itself concludes that "sequence-based approaches to protein-function prediction have proved to be very useful" (Skolnick at page 37), admitting that such methods have correctly assigned function in 50-70% of the cases, thus arguing against the conclusion drawn in the Action.

The Action next cites Doerks *et al.* (Trends in Genetics *14*:248-250, 1998) in support that sequence-to-function methods of assigning protein function are prone to errors due to partial annotation, multifunctionality and over prediction. However, Doerks *et al.* states that "utilization of family information and thus a more detailed characterization" should lead to "simplification of update procedures for the entire families if functional information becomes available for at least one member" (Doerks *et al.*, page 248, paragraph bridging columns 1 and 2, emphasis added). Applicants point out that transporters represent a well-studied protein family with a large amount of known functional information, exactly the situation that Doerks *et al.* suggests will "simplify" and "avoid the pitfalls" of

previous sequence-to-function methods of assigning protein function (Doerks *et al.*, page 248, columns 1 and 2). Thus, instead of supporting the Action's position against utility, Doerks *et al.* supports Applicants' position that the presently claimed sequences have a recognized substantial and credible utility.

The Action next cites Smith and Zhang (Nature Biotechnology *15*:1222-1223, 1997). First it should be noted that the Smith and Zhang article precedes the filing date of the provisional U.S. patent application no. 60/275,013, filed on March 12, 2001, to which the present application claims benefit by four years and may therefore no longer represents the state of the art years later. Aside from that fact, the Smith and Zhang article also states "the major problems associated with nearly all of the current automated annotation approaches are - paradoxically - minor database annotation inconsistencies (and a few outright errors)" (page 1222, second column, first paragraph, emphasis added). Thus, Smith and Zhang do not in fact seem to stand for the proposition that prediction of function based on homology is fraught with uncertainty, and thus also does not support the alleged lack of utility.

The Action also cites Brenner (Trends in Genetics *15*:132-133, 1999) as teaching that most homologs must have different molecular and cellular functions. However, this statement is based on the assumption that "if there are only 1000 superfamilies in nature, then most homologs must have different molecular and cellular functions (page 132, second column). Furthermore, Brenner suggests that one of the main problems in using homology to predict function is "an issue solvable by appropriate use of modern and accurate sequence comparison procedures" (page 132, second column), and in fact references an article by Altschul *et al.*, which is the basis for one of the "modern and accurate sequence comparison procedures" used by Appellants. Thus, the Brenner article also does not support the alleged lack of utility.

Finally, the Action cites Bork and Bairoch (Trends in Genetics *12*:425-427, 1996) as supporting the proposition that prediction of protein function from homology information is somewhat unpredictable. The question as to whether Bork's positions are generally supported by those of skill in the art was discussed above in the paragraph regarding the other Bork citations. It should also be noted that this article was published approximately 8 years ago and thus refers to errors or "traps" associated with earlier algorithms and technologies in a field that has undergone constant improvement. This publication identifies (Table 1) various areas in which incorrect information appears in sequence

databases. These "traps" include Synonyms - a single gene having a variety of names, Different gene-same name- when the same name is used to describe different genes, Spelling errors, Contamination- the unintentional inclusion of vector sequences, etc. and propagation of incorrect functional associations based on poorly analyzed homology. All of these issues can effect the accuracy of sequence base analysis, however all can be overcome by a more careful analysis as would be done by one of skill in the art. Automatic methods of sequence homology as identified by any algorithm is a staring point for consideration, and one of skill in the art can then through further analysis, structure-function analysis, etc. can and should then verify the associations. For example in addition to algorithm based sequence analysis the sequences of the present invention underwent careful analysis by a series of individuals of skill in the art, many highly qualified (1 B.S. and 3 Ph.D. level scientists). Clearly such highly skilled and careful analysis reduces the influence of such "traps". Furthermore, in the final section of this publication (page 427) it again becomes cleat that Bork and Bairoch do not discount the value of sequence analysis "we wish to point out that sequence database are the most useful tool in sequence analysis and the question should be how can one further improve their value". Thus clearly this publication represents a call to action to enhance the already high value of sequence analysis rather than an indictment of the utility of sequence based analysis. Therefore, as Bork and Bairoch identifies the high value of sequence based analysis it actually supports rather than refutes Applicants assertions regarding the utility of the present invention.

In summary, a careful reading of the cited "relevant literature" does not in fact support the concept that function cannot be based on sequence and structural similarity, in contrast many of the examples actually support the use of such methodologies while identifying several areas in which caution should be exercised. As stated previously these inaccuracies and potential pitfalls can be overcome by a more careful analysis by those of skill in the art. Automatic methods of sequence homology identification was only the staring point for consideration the sequences of the present invention underwent careful analysis by a series of individuals of skill in the art, many highly qualified (1 B.S. and 3 Ph.D. level scientists).

Furthermore, these articles are just examples of the few contrarian articles that the PTO has repeatedly attempted to use to deny the utility of nucleic acid sequences based on a small number of publications that call into doubt prediction of protein function from homology information and the usefulness of bioinformatic predictions. While there may not be a 100% consensus within the scientific

community regarding prediction of protein function from homology information, this is not unusual nor is it indicative of a general lack of consensus. A few rare exceptions do not a rule make.

The position that bioinformatic information is recognized to be of value by those of skill in the art is supported by the results of a recent search of the NCBI-NLM-NIH public scientific database "PubMed" using the term "bioinformatics" which resulted in 5,548 different scientific publications (these will not be provided to avoid burdening the USPTOs scanning group). If bioinformatic information is not useful in predicting protein function from structural homology information, why are so many publications reporting the results of its use? Clearly this suggest that those of skill in the art do recognize bioinformatic data as useful and valid.

A second form of evidence supporting the position that bioinformatic information is recognized to be of value by those of skill in the art is the fact that many scientists, corporations and institutions elect to allocate significant proportions of their limited resources for access to private bioinformatic systems and databases. Thus, it would appear obvious that those of skill in the art value and accept the results of bioinformatic analysis for they are willing to pay dearly for access to such information.

A third, an perhaps most persuasive, form of evidence supporting the position that bioinformatic information is recognized to be of value by those of skill in the art is the issuance of multiple US patents regarding bioinformatic prediction and methods for doing the same (see for example, U.S. Patent Nos. 6,229,911, 6,567,540, 6,615,141, 6,631,331, 6,651,008, 6,677,114, these patents will not be provided to avoid burdening the USPTOs scanning group). Of particular interest might be U.S. Patent No. 6,466,874, one of whose claims reads on "A method of identifying proteins as functionally linked, the method comprising comparing sequences to find homologous functional domains." Why would a U.S. Patent have issued on a method of carrying out an analysis that is without utility, because it is not accepted by those of skill in the art as a credible method of predicting function from structural homology information? This evidence convincingly indicates that even the USPTO recognizes the utility of bioinformatic prediction.

Appellants respectfully point out that, as discussed above, the legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be believable. Appellants submit that the overwhelming majority of those of skill in the relevant art would believe prediction of protein function from homology information and the usefulness of bioinformatic predictions to be powerful and useful tools. Clearly the several forms of evidence

presented, and certainly the issuance of U.S. Patents suggest that those of skill in the art recognize the utility of bioinformatic analysis and its credibility in assessing structure function relationships. Thus the vast majority of those of skill in the art (and the present Examiner) would believe that Appellants sequence encodes an EGF protein (specifically MEGF10) .

Rather, the question of utility is a straightforward one. As set forth by the Federal Circuit, "(t)he threshold of utility is not high: An invention is 'useful' under section 101 if it is capable of providing some identifiable benefit." *Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing *Brenner v. Manson*, 383 U.S. 519, 534 (1966)). Additionally, the Federal Circuit has stated that "(t)o violate § 101 the claimed device must be totally incapable of achieving a useful result." *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 (Fed. Cir. 1992), emphasis added. *Cross v. Iizuka* (224 USPQ 739 (Fed. Cir. 1985); "*Cross*") states "any utility of the claimed compounds is sufficient to satisfy 35 U.S.C. § 101". *Cross* at 748, emphasis added. Indeed, the Federal Circuit recently emphatically confirmed that "anything under the sun that is made by man" is patentable *(State Street Bank & Trust Co. v. Signature Financial Group Inc.*, 47 USPQ2d 1596, 1600 (Fed. Cir. 1998), citing the U.S. Supreme Court's decision in *Diamond vs. Chakrabarty*, 206 USPQ 193 (S.Ct. 1980)).

The legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be credible or believable. According to the Examination Guidelines for the Utility Requirement, if the applicant has asserted that the claimed invention is useful for any particular purpose (i.e., it has a "specific and substantial utility") and the assertion would be considered credible by a person of ordinary skill in the art, the Examiner should not impose a rejection based on lack of utility. Therefore, the present claims clearly meet the requirements of 35 U.S.C. § 101.

The Action suggests that Applicants knew of no specific use at the time the application was filed that would permit an immediate use by the public of a disclosed nucleic acid sequence. However, the specification details a number of uses for the presently claimed polynucleotide sequences, among these were, use in diagnostic assays such as forensic analysis (see, for example, the specification at page 3, line 9 and page 12, line 4) in identification of protein coding sequence and identification of exon splice junctions (see, for example, the specification at page 3, line 5, and page 11, line 21), in mapping the sequences to a specific region of a human chromosome (see, for example, the specification at page 3,

line 3; page 11, line 19 and particularly page 17, lines 15-18), and in assessing gene expression patterns, particularly using a high throughput "chip" format (see, for example, the specification at page 6, line 3 through page 8).

In the Action the position that the argument that the polymorphisms described in the specification have patentable utility is also deemed to be non-persuasive. This is allegedly because the asserted utilities of the present nucleic acid sequences and their identified polymorphisms in forensic analysis, human population biology, or paternity identification are not specific or substantial.

Naturally occurring genetic polymorphisms such as those described in the present specification are both the basis of, and critical to, *inter alia*, forensic genetic analysis and genetic analysis intended to resolve issues of identity and paternity. Therefore, Applicants find this position difficult to comprehend, given that the results of identity and paternal analysis often have great emotional and substantial economic impact. This does not sound like a throw away utility, rather it sounds like a very substantial and real world utility. What could be more substantial and real world than the loss of an individual's freedom through incarceration and in some cases even the loss of life through execution? Yet forensic analysis based on identified polymorphisms is often used to convict or acquit in many cases. Both paternal and forensic genetic analysis is based on the use of identified polymorphisms. This is a well known and generally accepted by those of skill in the art, who would readily recognize the utility and value of any identified polymorphism. Without identified polymorphisms, one would not be able to carry out such forensic or paternal analyses. The present application has identified just such essential polymorphisms within the sequences of the present invention which identify as EGF protein (MEGF10).

As such polymorphisms are the basis for forensic analysis, paternity identification and population biology studies, which are undoubtedly "real world" utilities, the present sequences must in themselves be useful. In and of themselves each of these polymorphisms, including the silent ones, has significant and specific utility, the specificity of this utility is only amplified by the presence of so many polymorphisms that can arise in various combinations. It is also important to note that the presence of more useful polymorphic markers for such analysis would not mean that the present sequences lack utility.

Applicants respectfully point out that those of skill in the art would readily recognize that the presently described polymorphisms, exactly as they were described in the specification as originally

filed, are useful in forensic analysis, population biology and paternity analysis to specifically identify individual members of the human population based on the presence or absence of the described polymorphism. Simply because the use of these polymorphic markers will necessarily provide additional information on the percentage of particular subpopulations that contain one or more of these polymorphic markers does not mean that "additional research" is needed in order for these markers as they are presently described in the instant specification to be of use to forensic science. Without further experimentation those of skill in the art would recognize the utility of the identified polymorphisms and how the asserted markers can distinguish 50% of the population in the worst case scenario. Thus the presence or the absence of a particular specific polymorphism is sufficient for use in the proposed utilities. Applicants provide the following detailed explanation. Those of skill in the art would recognize that in the worst case, least useful situation, a marker would be present in half of a population and absent from the other half. Therefore the probability of an individual having such a marker would be 1 in 2 or 50%. Using the forensic analysis scenario for example, the analysis will have removed 50% of the possible suspects from the list, as either the suspect has the identified polymorphism or not. However, if a polymorphism were present in only say 10% of the population, the probability of an individual having such a polymorphic marker would be 1 in 10 (10%) and 90% of suspects could be eliminated from investigation or prosecution based on the presence or absence of the polymorphism. Clearly eliminating 90% of the suspects is better than eliminating 50% of the suspects. That said, eliminating 50% or half of the suspects on a list is without question very useful to any investigator. To reiterate, using the polymorphic markers as described in the specification as originally field will definitely distinguish members of a population from one another. In the worst case scenario, each of these markers are useful to distinguish 50% of the population (in other words, the marker being present in half of the population). The ability to eliminate 50% of the population from a forensic analysis clearly is a real world, practical utility. Therefore, any allegation that the use of the presently described polymorphic markers is only potentially useful would be completely without merit, and would not support the alleged lack of utility.

Perhaps the Examiner is assuming that since any human nucleic acid sequence that contains a naturally occurring polymorphism can be used in forensic analysis, in human paternity determinations or human population migration determinations, such utilities are generic and therefore lack substantial and specific utility. First, Applicants submit that until a specific polymorphic marker is actually

described it has very limited utility in forensic analysis. Put another way, simply because there is a possibility, even a significant likelihood, that a particular nucleic acid sequence will contain a polymorphism and thus be useful in forensic analysis, until such a specific polymorphism is actually identified and described, such a likelihood is meaningless. The present case contains identified polymorphisms that occur in a novel human EGF protein. The Examiner is perhaps attempting to use the information presented for the first time by Applicants in the instant specification as hindsight verification that the presently claimed sequence would be expected to have polymorphic markers. Such a hindsight analysis based on Applicants discovery would not be proper.

Alternatively, the assumption that since any sequence containing a naturally occurring polymorphism can be used such utilities are generic and therefore lack substantial and specific utility may represent a confusion between the requirement for a specific utility, which is the proper standard for utility under 35 U.S.C. § 101, with a requirement for a unique utility. The relevant case law cited by Applicants makes it abundantly clear that the presence of other or even more useful polymorphic markers for forensic analysis does **not** mean that the present sequences lack a specific utility. As clearly stated by the Federal Circuit in *Carl Zeiss Stiftung v. Renishaw PLC*, 20 USPQ2d 1101 (Fed. Cir. 1991; "*Carl Zeiss*"):

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding a lack of utility." *Envirotech Corp. v. Al George, Inc.*, 221 USPQ 473, 480 (Fed. Cir. 1984)

Importantly, the holding in the *Carl Zeiss* case is mandatory legal authority that essentially controls the outcome of the present appeal. This case, and particularly the cited quote, directly rebuts any such argument. Furthermore, the requirement for a unique utility is clearly not the standard adopted by the Patent and Trademark Office. If every invention were required to have a unique utility, the Patent and Trademark Office would no longer be issuing patents on batteries, automobile tires, golf balls, golf clubs, and treatments for a variety of human diseases, such as cancer and bacterial or viral infections, just to name a few particular examples, because examples of each of these have already been described and patented. All batteries have the exact same utility - specifically, to provide power. All automobile

tires have the exact same utility - specifically, for use on automobiles. All golf balls and golf clubs have the exact same utility - specifically, use in the game of golf. All cancer treatments have the exact same utility - specifically, to treat cancer. All anti-infectious agents have the exact same broader utility - specifically, to treat infections. However, only the briefest perusal of virtually any issue of the Official Gazette provides numerous examples of patents being granted on each of the above compositions every week. Furthermore, if a composition needed to be unique to be patented, the entire class and subclass system would be an effort in futility, as the class and subclass system serves solely to group such common inventions, which would not be required if each invention needed to have a unique utility. Thus, the present sequence clearly meets the requirements of 35 U.S.C. § 101.

Although the above discussion is believed to be dispositive of the utility issue, Applicants would like to further direct the Examiner's attention to the parts of the specification that describe the use of sequences in a gene chip format to provide a high throughput analysis of the relevant cellular "transcriptome", including assessing temporal and tissue specific gene expression patterns, particularly using a high throughput "chip" format (specification at or about page 6, line 3 through page 8).

Evidence of the "real world" substantial utility of the present invention is further provided by the fact that there is an entire industry established based on the use of gene sequences or fragments thereof in a gene chip format. Perhaps the most notable gene chip company is Affymetrix. However, there are many companies which have, at one time or another, concentrated on the use of gene sequences or fragments, in gene chip and non-gene chip formats, for example: Gene Logic, ABI-Perkin-Elmer, HySeq and Incyte. In addition, one such company, Rosetta Inpharmatics, was viewed to have such "real world" value that it was acquired by large pharmaceutical company, Merck & Co., for substantial sums of money (net equity value of the transaction was $620 million). The "real world" substantial industrial utility of gene sequences or fragments would, therefore, appear to be widespread and well established. Clearly, persons of skill in the art, as well as venture capitalists and investors, readily recognize the utility, both scientific and commercial, of genomic data in general, and specifically human genomic data. Billions of dollars have been invested in the human genome project, resulting in useful genomic data (see, *e.g.*, Venter *et al.*, 2001, Science *291*:1304, presented as **Exhibit A**). The results have been a stunning success as the utility of human genomic data has been widely recognized as a great gift to humanity (see, *e.g.*, Jasny and Kennedy, 2001, Science *291*:1153, presented as **Exhibit B**). Clearly, the usefulness of human genomic data, such as the presently claimed nucleic acid molecules,

is substantial and credible (worthy of billions of dollars and the creation of numerous companies focused on such information) and well-established (the utility of human genomic information has been clearly understood for many years). The sequences of the present invention have particularly specific utility in DNA gene chip based analysis as they have been identified to contain several coding region single nucleotide polymorphisms (cSNPs), thus increasing their utility in DNA gene chip based analysis.

DNA chips clearly have utility, as evidenced by hundreds of issued U.S. Patents, as exemplified by U.S. Patent Nos. 5,445,934, 5,556,752, 5,744,305, 5,837,832, 6,156,501 and 6,261,776 (**Exhibits C-H**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy). Accordingly, the present sequence has a specific utility in such DNA chip applications. Clearly, compositions that enhance the utility of such DNA chips, like the present sequences, which encode a human EGF protein, have identified polymorphisms and a characterized tissue expression pattern, must have utility. The sequences of the present invention which encode the human EGF protein ,MEGF10, provide specific markers for a human genome (see also chromosome mapping discussion below and information provided in the specification at page 17, lines 16-18 that indicate that this protein is encoded on human chromosome 5). Thus, those skilled in the art would instantly recognize that the sequences of the present invention would be an ideal, novel candidate for assessing gene expression using, for example, DNA chips, as the specification details. Accordingly, the present sequence has a specific utility in such DNA chip applications. Clearly, compositions that enhance the utility of such DNA chips, such as the presently claimed nucleotide sequence encoding human EGF protein, MEGF10, must also be useful. Additional support for the position that those of skill in the art recognize the utility of the sequences of the present invention encoding MEGF10 on gene chips, is the fact that such chips are available from several sources. For Example, at the Keck Biotechnology Resource Laboratory at the Yale University School of Medicine, presented as **Exhibit I** is the contact page posted on the internet (http://keck.med.yale.edu/DNAarrays/contact_arrays.html) and a partial listing of the genes available on arrays at this location (http://keck.med.yale.edu/DNAarrays/OHU21K.gal), MEGF10 appears on page 12 of the partial listing and is present at chip position Block 2, Row 8, Column 18. MEGF10 is also represented on a microarray chip available at the the Flanders Interuniversity Institute for for Biotechnology (VIB), presented as **Exhibit J** is the contact page posted on the internet (www.microarrays.be/Home.htm) and a partial listing of the genes available on arrays at this location

(www.microarrays.be/download/arrays/Hs/archived_CIUS/Hs4/VIB_Human_21K_I.txt), MEGF10 appears on page 13 of the partial listing at position 748 of chip HS4. Thus, clearly those of skill in the art must recognize the utility of using the sequences (and fragments thereof) of the present invention on gene chips as they make access to them available for a fee, indicating that they believe others of skill in the art will also recognize the value and be willing to pay for access.

The Examiner is further requested to reconsider that, given the huge expense of the drug discovery process, even negative information obtained using these specific markers of expression of a human EGF protein provides very specific markers for the human genome and have great "real world" practical utility. Knowing that a given gene is not expressed in medically relevant tissue provides an informative finding of great value to industry by allowing for the more efficient deployment of expensive drug discovery resources. Such practical considerations are equally applicable to the scientific community in general, in that time and resources are not wasted chasing what are essentially scientific dead-ends (from the perspective of medical relevance). Clearly, compositions that enhance the utility of DNA gene chips, such as the presently claimed sequences encoding a human EGF protein, must in themselves be useful. Moreover, the presently described human EGF protein sequences provide uniquely specific sequence resources for identifying and quantifying full length transcripts that were encoded by the corresponding human genomic locus. Accordingly, there can be no question that the described sequences provide an exquisitely specific utility for analyzing gene expression. Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Further evidence of utility of the presently claimed polynucleotide, although only one is needed to meet the requirements of 35 U.S.C. § 101 (*Raytheon v. Roper*, 220 USPQ 592 (Fed. Cir. 1983); *In re Gottlieb*, 140 USPQ 665 (CCPA 1964); *In re Malachowski*, 189 USPQ 432 (CCPA 1976); *Hoffman v. Klaus*, 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988)), is the specific utility the present nucleotide sequence has in determining the genomic structure of the corresponding human chromosome, for example mapping the protein encoding regions as described in the specification. Clearly, the present polynucleotide provides exquisite specificity in localizing the specific region of the human chromosome 5 containing the gene encoding the given polynucleotide (MEGF10), a utility not shared by virtually any other nucleic acid sequence. In fact, it is this specificity that makes this particular sequence so useful. Early gene mapping techniques relied on methods such as Giemsa staining to identify regions of chromosomes. However, such techniques produced genetic maps with a resolution of only 5 to 10

megabases, far too low to be of much help in identifying specific genes involved in disease. The skilled artisan readily appreciates the significant benefit afforded by markers that map a specific locus of the human genome, such as the present nucleic acid sequence.

The Action discounts Appellants' assertion regarding the use of the presently claimed polynucleotides for gene mapping and determining chromosome structure again based on the position that such a use would allegedly be generic and therefore fail to represent a specific and substantial utility. However, as only a minor percentage of the genome actually encodes exons, which in turn encode amino acid sequences, the presently claimed polynucleotide sequence provides biologically validated empirical data (*e.g.,* showing which sequences are transcribed, spliced, and polyadenylated) that *specifically* defines that portion of the corresponding genomic locus that actually encodes exon sequence. Equally significant is that the claimed polynucleotide sequence defines how the encoded exons are actually spliced together to produce an active transcript (*i.e.*, the described sequences are useful for functionally defining exon splice-junctions). The Applicants respectfully submit that the practical scientific value of expressed, spliced, and polyadenylated mRNA sequences is readily apparent to those skilled in the relevant biological and biochemical arts. For further evidence supporting the Applicants' position, the Board is requested to review, for example, section 3 of Venter *et al.* (*supra at* pp. 1317-1321, including Fig. 11 at pp.1324-1325), which demonstrates the significance of expressed sequence information in the structural analysis of genomic data. The presently claimed polynucleotide sequence defines a biologically validated sequence that provides a unique and specific resource for mapping the genome essentially as described in the Venter *et al.* article.

The Action states (at page 10) that the mapping description provided in the specification (at page 17, lines 16-18) was not precise enough and takes the position that using the information provided in the specification "substantial further research would be required for the skilled artisan to determine this precise position. Apparently the Examiner views the overlay of SEQ ID NO:1 onto the known human genome sequence "substantial further research" as that is how the evidence provided in **Exhibit K** was obtained. This evidence supports Applicants' assertions of the specific utility of the sequences of the present invention in localizing the specific region of the human chromosome and identification of functionally active intron/exon splice junctions. **Exhibit K** is the result of overlaying the sequence of SEQ ID NO:1 of the present invention and the identified human genomic sequence. By doing this one is readily able to identify the portions of the genome that encode the present invention. If these regions

of the genome are non-contiguous, this is indicative of individual exons. The results of such an analysis indicates that the sequence of the present invention is encoded by 24 exons spread non-contiguously along a region of human chromosome 5, (as stated in the specification as filed on page 17, lines 16-18) at approximately 5q33 (which is also contained within clone, AC008682.6). Thus clearly one would not simply be able to identify the 24 distinct protein encoding exons that make up the sequence of the present intention from within the large genomic sequence. Nor, would one be able to map the protein encoding regions identified specifically by the sequences of the present invention without knowing exactly what those specific sequences were. Additionally, it should be noted that the human EGF protein, MEGF10 gene is now recognized to map to the same region of human chromosome 5 (5q33). This further supports Applicant's position that the sequences of the present invention encodes a human EGF protein, MEGF10 and was described in the specification.

In addition, among other things the mapping of the relatively few expressed human genes to a particular chromosome has long been a recognized method of identifying a genes associated with particular diseases. Furthermore, the mapping of the human chromosome is a project of such widely recognized importance by those of skill in the art and even lay people, that both the US government and private corporations have dedicated millions of dollars to such a project. One is thus forced to ask, if the mapping of human chromosomes is a throw away utility then why has the US government spent so many taxpayer dollars on this project?

The Action's repeated position that this utility, like the use of these specific sequences on DNA chips or the described polymorphisms in forensic analysis, is that since other molecules can be used to map the human chromosome or on DNA chips or in forensic analysis, these utilities are not specific or substantial. As described previously above, Appellants once again point out that these arguments are completely rebuffed by the Federal Circuit's holding in *Carl Zeiss*, *supra* ("[A]n invention need not be the best or only way to accomplish a certain result"). Furthermore, the argument that just because there are other objects having the same utility, that utility has been rendered generic and therefore invalid begs the question, previously presented, that don't all golf balls and tires have the same utility of other golf balls or tires, i.e. they can be used as golf balls or tires respectively and yet these items are readily considered to have patentable utility.

It has been clearly established that a statement of utility in a specification must be accepted absent reasons why one skilled in the art would have reason to doubt the objective truth of such

statement. *In re Langer*, 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974; "*Langer*"); *In re Marzocchi*, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971). As clearly set forth in *Langer*:

> As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope.

Langer at 297, emphasis in original. As set forth in the MPEP, "Office personnel must provide evidence sufficient to show that the statement of asserted utility would be considered 'false' by a person of ordinary skill in the art" (MPEP, Eighth Edition at 2100-40, emphasis added). Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Finally, with full recognition of the fact that all patent applications are examined on their own merits and that the prosecution of one patent does not effect the prosecution of another patent, *In re Wertheim,* 541 F.2d 257, 264, 191 USPQ 90, 97 (CCPA 1976), however the issue at hand in one of whether the fact that patents have issued recognizing the utility of a class of molecules does this confers a statutory precedent of patentability to a broad class of compositions. Thus, there remains a lingering issue regarding due process and equitable treatment under the law. While Applicants are well aware of the new Utility Guidelines set forth by the USPTO, Applicants respectfully point out that the current rules and regulations regarding the examination of patent applications is and always has been the patent laws as set forth in 35 U.S.C. and the patent rules as set forth in 37 C.F.R., not the Manual of Patent Examination Procedure or particular guidelines for patent examination set forth by the USPTO. Furthermore, it is the job of the judiciary, not the USPTO, to interpret these laws and rules. Applicants are unaware of any significant recent changes in either 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit that is in keeping with the new Utility Guidelines set forth by the USPTO. This is underscored by numerous patents that have been issued over the years that claim nucleic acid fragments that do not comply with the new Utility Guidelines. As examples of such issued U.S. Patents, the Examiner is invited to review U.S. Patent Nos. 5,817,479, 5,654,173, and 5,552,281 (each of which claims short polynucleotides; **Exhibits L-**

N; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy), and recently issued U.S. Patent No. 6,340,583 (which includes no working examples; **Exhibit O**; copies of issued U.S. Patents not provided pursuant to current United States Patent and Trademark Office policy), none of which contain examples of the "real-world" utilities that the Examiner appears to desire. As issued U.S. Patents are presumed to meet all of the requirements for patentability, including 35 U.S.C. §§ 101 and 112, first paragraph (see Section IV, below), Applicants submit that the present polynucleotides must also meet the requirements of 35 U.S.C. § 101. While Applicants agree that each application is examined on its own merits, Applicants are unaware of any changes to 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit, since the issuance of these patents that render the subject matter claimed in these patents, which is similar to the subject matter in question in the present application, as suddenly non-statutory or failing to meet the requirements of 35 U.S.C. § 101. Thus, holding Appellants invention to a different standard of utility appears inconsistent and inequitable, such a judgement being arbitrary and capricious, a violation of due process and equal protection under the law and cannot be maintained.

In light of the evidence presented herewith and for the many compelling reasons described above, it is clear that the present invention clearly has utilities that are specific, substantial and credible. Therefore, Applicants submit that the rejection of the pending claims under 35 U.S.C. § 101 has been avoided and respectfully request withdrawal of the pending rejection of claims under 35 U.S.C. § 101.

IV. Rejection of Claims Under 35 U.S.C. § 112, First Paragraph

The Action also rejects claims under 35 U.S.C. § 112, first paragraph, since allegedly one skilled in the art would not know how to use the invention, as the invention allegedly is not supported by a specific, substantial, and credible utility or a well-established utility. Applicants respectfully traverse.

Applicants submit that as the present invention has been shown to have "a specific, substantial, and credible utility", as detailed in section II above, the rejection under 35 U.S.C. § 112, first paragraph, has been avoided. Applicants therefore request that the rejection of the pending claims under 35 U.S.C. § 112, first paragraph, be withdrawn.

V. Rejection of Claims Under 35 U.S.C. § 112, First Paragraph

The Action next rejects claim 1-4 (although Applicants believe that the Examiner intends only claim 1) under 35 U.S.C. § 112, first paragraph, as allegedly containing subject matter that was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. Applicants respectfully traverse.

35 U.S.C. § 112, first paragraph, requires that the specification contain a written description of the invention. The Federal Circuit in *Vas-Cath Inc. v. Mahurkar* (19 USPQ2d 1111 (Fed. Cir. 1991); "*Vas-Cath*") held that an "applicant must convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention*." *Vas-Cath*, at 1117, emphasis in original. However, it is important to note that the above finding uses the terms reasonable clarity to those skilled in the art. Further, the Federal Circuit in *In re Gosteli* (10 USPQ2d 1614 (Fed. Cir. 1989); "*Gosteli*") held:

> Although [the applicant] does not have to describe exactly the subject matter claimed, . . . the description must clearly allow persons of ordinary skill in the art to recognize that [he or she] invented what is claimed.

Gosteli at 1618, emphasis added. Additionally, *Utter v. Hiraga* (6 USPQ2d 1709 (Fed. Cir. 1988); "*Utter*"), held "(a) specification may, within the meaning of 35 U.S.C. § 112 ¶1, contain a written description of a broadly claimed invention without describing all species that claim encompasses" (*Utter*, at 1714). Therefore, all Applicants must do to comply with 35 U.S.C. § 112, first paragraph, is to convey the invention with reasonable clarity to the skilled artisan.

The Action, at page 12, states "With the exception of **SEQ ID NO: 1**, the skilled artisan cannot envision the detailed chemical structure of the encompassed polypeptides, and therefore conception is not achieved until reduction to practice has occurred". Applicants in no way agree and respectfully point out that the Examiner is clearly using an **improper** standard for compliance with the written description requirement of 35 U.S.C. § 112, first paragraph. In the PTO Guidelines (66 Fed. Reg. at 1106), the PTO has determined that the written description requirement can be met by "show[ing] that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics . . . *i.e.*, complete or partial structure, other physical and/or chemical properties,

functional characteristics when coupled with a known or disclosed correlation between function and structure, **or** some combination of such characteristics" (66 Fed. Reg. at 1106, emphasis added). The Federal Circuit has recently confirmed this aspect of the PTO Guidelines, wherein this exact quote was reproduced (*Enzo Biochem, Inc. v. Gen-Probe, Inc. et al.* (296 F.3d 1316, 63 USPQ2d 1609 (Fed. Cir. 2002)). Taking the exact statement from the PTO Guidelines clause by clause, the written description requirement for a claimed genus may be satisfied through disclosure of sufficiently detailed, relevant identifying characteristics, which are defined as: (a) complete or partial structure; (b) other physical and/or chemical properties; (c) functional characteristics when coupled with a known or disclosed correlation between function and structure; **or** (d) some combination of such characteristics. In other words, the written description requirement is satisfied by (a), (b), (c) **or** (d). Clause (a) states that the written description requirement may be satisfied by the disclosure of structure. The Federal Circuit has held that an adequate description of a chemical genus "requires a precise definition, such as by structure, formula, chemical name or physical properties" sufficient to distinguish the genus from other materials. *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993; "*Fiers*"). *Fiers* goes on to hold that the "application satisfies the written description requirement since it sets forth the . . . nucleotide sequence" (*Fiers* at 1607). In other words, provision of a structure and formula - the nucleotide sequence - renders the application in compliance with 35 U.S.C. § 112, first paragraph. Therefore, claim 1 clearly meets the written description requirement of 35 U.S.C. § 112, first paragraph.

More recently, the standard for complying with the written description requirement in claims involving chemical materials has been explicitly set forth by the Federal Circuit:

> In claims involving chemical materials, generic formulae usually indicate with specificity what the generic claims encompass. One skilled in the art can distinguish such a formula from others and can identify many of the species that the claims encompass. Accordingly, such a formula is normally an adequate description of the claimed genus. *Univ. of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997).

Thus, a claim describing a genus of nucleic acids by structure, formula, chemical name or physical properties sufficient to allow one of ordinary skill in the art to distinguish the genus from other materials

meets the written description requirement of 35 U.S.C. § 112, first paragraph. As further elaborated by the Federal Circuit in *Univ. of California v. Eli Lilly and Co.*:

> In claims to genetic material ... a generic statement such as 'vertebrate insulin cDNA' or 'mammalian insulin cDNA', without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function. It does not specifically define any of the genes that fall within its definition. It does not define any structural features commonly possessed by members of the genus that distinguish them from others. One skilled in the art cannot, as one can do with a fully described genus, visualize or recognize the identity of members of the genus. (Emphasis added)

Thus, as opposed to the situation set forth in *Univ. of California v. Eli Lilly and Co.* and *Fiers*, the nucleic acid sequences of the present invention are not distinguished on the basis of function (as seemingly required by the Action), or a method of isolation, but in fact are distinguished by structural features - a chemical formula, *i.e.*, the *sequence itself.*

Using the nucleic acid sequences of the present invention (as set forth in the Sequence Listing), the skilled artisan would readily be able to **distinguish** the claimed nucleic acids from other materials on the basis of the specific **structural** description provided. It appears that the Examiner is asserting that those of skill in the art would be unable to identify 24 contiguous bases of SEQ ID NO:1, Applicants find this assertion to lack credibility and invite the Examiner to submit evidence to this effect.

Applicants respectfully point out that this **structural** characterization is **all that is required** of claim 1 to meet the written description requirement of 35 U.S.C. § 112, first paragraph. Polynucleotides comprising at least 24 contiguous bases of nucleotide sequence first disclosed in SEQ ID NO:1 are within the genus of the instant claims, while those that lack this structural feature lie outside the genus. The claimed genus of polynucleotides is clearly defined in **structural** terms, which is **all that is required** in order to meet the written description requirement of 35 U.S.C. § 112, first paragraph. Claim 1 thus meets the written description requirement.

However, as claim 1 has been cancelled without prejudice and without disclaimer the rejection of claim 1 under 35 U.S.C. § 112, first paragraph, has been rendered moot and should be withdrawn.

VI. Rejection Under 35 U.S.C. § 112, Second Paragraph

The Action rejects claim 2 under 35 U.S.C. § 112, second paragraph, as allegedly being indefinite for failing to particularly point out and distinctly claim the invention. Claim 2 stands rejected because the phrase "stringent conditions" is alleged to be indefinite. Although Applicants believe that this claim as originally filed sufficiently points out and distinctly claims the invention, as Applicants have cancelled claim 2 without prejudice and without disclaimer this issue has been rendered moot. Accordingly, the Examiner is respectfully requested to withdraw the pending rejection of claim 2 under 35 U.S.C. § 112, second paragraph.

VII. Conclusion

The present document is a full and complete response to the Action. In conclusion, Applicants submit that, in light of the foregoing remarks, the present case is in condition for allowance, and such favorable action is respectfully requested. Should Examiner Nichols have any questions or comments, or believe that certain amendments of the claims might serve to improve their clarity, a telephone call to the undersigned Applicants' representative is earnestly solicited.

Respectfully submitted,

March 11, 2004
Date

Lance K. Ishimoto Reg. No. 41,866
Attorney for Applicants

LEXICON GENETICS INCORPORATED
(281) 863-3333

Customer # 24231

EXHIBIT A

The Sequence of the Human Genome

J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]
Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5]
Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9]
Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1]
Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1]
Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1]
Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1]
Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1]
Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1]
Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1]
Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]
Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1]
Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12]
Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1]
Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1]
Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1]
Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1]
Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1]
Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1]
Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1]
Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1]
Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1]
Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1]
Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1]
Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1]
Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1]
Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1]
Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1]
Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1]
Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1]
Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1]
Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1]
Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1]
Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1]
Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1]
Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1]
Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1]
Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1]
Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1]
Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1]
Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1]
Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1]
Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1]
Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (*17*, *18*). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (*19*).

In 1997, Weber and Myers (*20*) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (*21*). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (*22*), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (*23*). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (*24*). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (*25*). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (*26*–*28*). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (*29*), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (*30*) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

1 Sources of DNA and Sequencing Methods

Summary. This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genome sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (*31*) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (*32*).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was

collected, as well as five specimens of semen, collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at www.sciencemag.org/cgi/content/291/5507/1304/DC1). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (*33*).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (*34*).

Current sequencing protocols are based on the dideoxy sequencing method (*35*), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (*36*).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (*37*). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (*26*). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

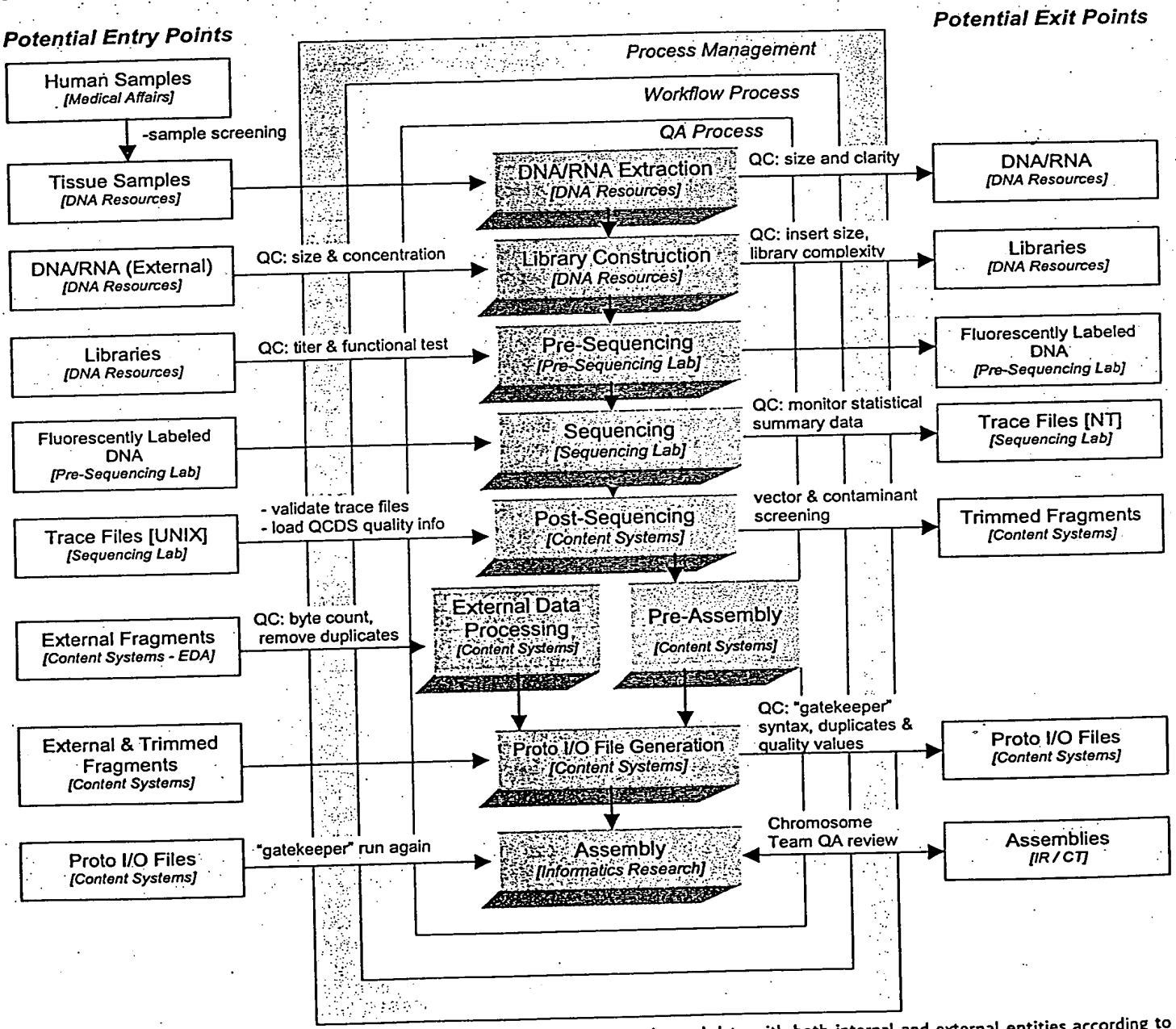


Fig. 2. Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange samples and data with both internal and external entities according to defined quality guidelines. Manufacturing pipeline processes, products, quality control measures, and responsible parties are indicated and are described further in the text.

and provide a comparison to the public genome sequence, which was reconstructed largely by an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff," and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (PFP) and is primarily derived from BAC clones (*30*). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (*38*), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Seqences division of GenBank (*39*), filtered at 200 bp at 98%; and (iii) the nonredundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (*18*).

2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (*40*). Furthermore, BAC location

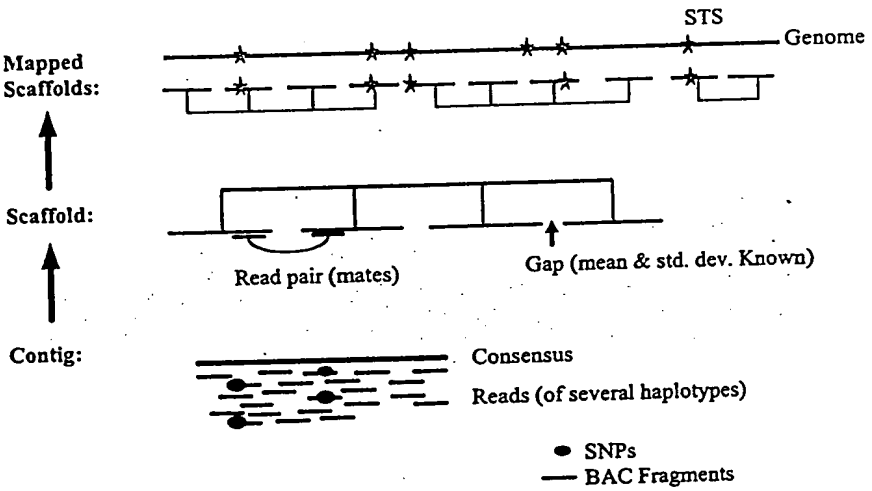


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that at least 2.2% of the BACs contained sequence data that were not part of the given BAC (*41*), possibly as a result of sample-tracking errors

Table 2. GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

(see below). In short, we performed a true, al initio whole-genome assembly in which w took the expedient of deriving additional se quence coverage, but not mate pairs, assembl bactigs, or genome locality, from some exte nally generated data.

In the compartmentalized shotgun assembl (CSA), Celera and PFP data were partitione into the largest possible chromosomal segment or "components" that could be determined wit confidence, and then shotgun assembly was ap plied to each partitioned subset wherein th bactig data were again shredded into faux read to ensure an independent ab initio assembly o the component. By subsetting the data in thi way, the overall computational effort was re duced and the effect of interchromosomal dupl cations was ameliorated. This also resulted in reconstruction of the genome that was relativel independent of the whole-genome assembly re sults so that the two assemblies could be con pared for consistency. The quality of the part tioning into components was crucial so th different genome regions were not mixed t gether. We constructed components from (i) th longest scaffolds of the sequence from ea BAC and (ii) assembled scaffolds of data uniqu to Celera's data set. The BAC assemblies we obtained by a combining assembler that used th bactigs and the 5× Celera data mapped to tho bactigs as input. This effort was undertaken an interim step solely because the more accura and complete the scaffold for a given sequen stretch, the more accurately one can tile the scaffolds into contiguous components on t basis of sequence overlap and mate-pair infor mation. We further visually inspected and c rated the scaffold tiling of the components further increase its accuracy. For the final CS assembly, all but the partitioning was ignore and an independent, ab initio reconstruction the sequence in each component was obtain by applying our whole-genome assembly alg rithm to the partitioned, relevant Celera data t the shredded, faux reads of the partitioned, r evant bactig data.

2.3 Whole-genome assembly

The algorithms used for whole-genome sembly (WGA) of the human genome w enhancements to those used to produce sequence of the *Drosophila* genome repor in detail in (*28*).

The WGA assembler consists of a pipe composed of five principal stages: Scree Overlapper, Unitigger, Scaffolder, and Re Resolver, respectively. The Screener fi and marks all microsatellite repeats with than a 6-bp element, and screens out known interspersed repeat elements, inc ing Alu, Line, and ribosomal DNA. Mu regions get searched for overlaps, whe screened regions do not get searched, but be part of an overlap that involves unscree matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in-10^{17} event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (*42*). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6× simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in 10^{10}, assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than 10^{-7} based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (*43*). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

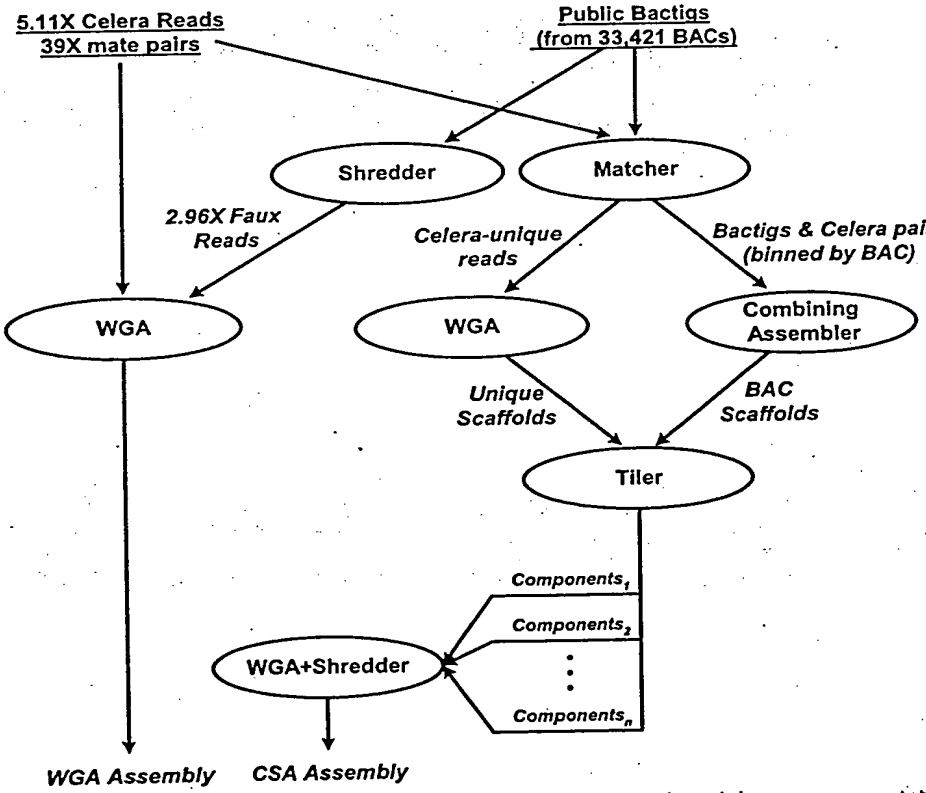


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

some 22, all stones were placed correctly.

The final method of resolving gaps is to fill them with assembled BAC data that cover the gap. We call this external gap "walking." We did not include the very aggressive "Pebbles" substage described in our *Drosophila* work, which made enough mistakes so as to produce repeat reconstructions for long interspersed elements whose quality was only 99.62% correct. We decided that for the human genome it was philosophically better not to introduce a step that was certain to produce less than 99.99% accuracy. The cost was a somewhat larger number of gaps of somewhat larger size.

At the final stage of the assembly process, and also at several intermediate points, a consensus sequence of every contig is produced. Our algorithm is driven by the principle of maximum parsimony, with quality-value–weighted measures for evaluating each base. The net effect is a Bayesian estimate of the correct base to report at each position. Consensus generation uses Celera data whenever it is present. In the event that no Celera data cover a given region, the BAC data sequence is used.

A key element of achieving a WGA of the human genome was to parallelize the Overlapper and the central consensus sequence–constructing subroutines. In addition, memory was a real issue—a straightforward application of the software we had built for *Drosophila* would have required a computer with a 600-gigabyte RAM. By making the Overlapper and Unitigger incremental, we were able to achieve the same computation with a maximum of instantaneous usage of 28 gigabytes of RAM. Moreover, the incremental nature of the first three stages allowed us to continually update the state of this part of the computation as data were delivered and then perform a 7-day run to complete Scaffolding and Repeat Resolution whenever desired. For our assembly operations, the total compute infrastructure consists of 10 four-processor SMPs with 4 gigabytes of memory per cluster (Compaq's ES40, Regatta) and a 16-processor NUMA machine with 64 gigabytes of memory (Compaq's GS160, Wildfire). The total compute for a run of the assembler was roughly 20,000 CPU hours.

The assembly of Celera's data, together with the shredded bactig data, produced a set of scaffolds totaling 2.848 Gbp in span and consisting of 2.586 Gbp of sequence. The chaff, or set of reads not incorporated in the assembly, numbered 11.27 million (26%), which is consistent with our experience for *Drosophila*. More than 84% of the genome was covered by scaffolds >100 kbp long, and these averaged 91% sequence and 9% gaps with a total of 2.297 Gbp of sequence. There were a total of 93,857 gaps among the 1637 scaffolds >100 kbp. The average scaffold size was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp, where the distribution of each was essentially exponential More than 50% of all gaps were less than 50(bp long, >62% of all gaps were less than 1 kbp long, and no gap was >100 kbp long. Similarly, more than 65% of the sequence is in contigs >30 kbp, more than 31% is in contigs >100 kbp, and the largest contig was 1.22 Mbp long. Table 3 gives detailed summary statistics for the structure of this assembly with a direct comparison to the compartmentalized shotgun assembly.

2.4 Compartmentalized shotgun assembly

In addition to the WGA approach, we pursued a localized assembly approach that was intended to subdivide the genome into segments, each of which could be shotgun assembled individually. We expected that this would help in resolution of large interchromosomal duplications and improve the statistics for calculating U-unitigs. The compartmentalized assembly process involved clustering Celera reads and bactigs into large, multiple megabase regions of the genome, and then running the WGA assembler on the Celera data and shredded, faux reads obtained from the bactig data.

The first phase of the CSA strategy was to separate Celera reads into those that matched the BAC contigs for a particular PFP BAC entry, and those that did not match any public data. Such matches must be guaranteed to

Table 3. Scaffold statistics for whole-genome and compartmentalized shotgun assemblies.

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
		Compartmentalized shotgun assembly			
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
		Whole-genome assembly			
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

properly place a Celera read, so all reads were first masked against a library of common repetitive elements, and only matches of at least 40 bp to unmasked portions of the read constituted a hit. Of Celera's 27.27 million reads, 20.76 million matched a bactig and another 0.62 million reads, which did not have any matches, were nonetheless identified as belonging in the region of the bactig's BAC because their mate matched the bactig. Of the remaining reads, 2.92 million were completely screened out and so could not be matched, but the other 2.97 million reads had unmasked sequence totaling 1.189 Gbp that were not found in the GenBank data set. Because the Celera data are 5.11× redundant, we estimate that 240 Mbp of unique Celera sequence is not in the GenBank data set.

In the next step of the CSA process, a combining assembler took the relevant 5× Celera reads and bactigs for a BAC entry, and produced an assembly of the combined data for that locale. These high-quality sequence reconstructions were a transient result whose utility was simply to provide more reliable information for the purposes of their tiling into sets of overlapping and adjacent scaffold sequences in the next step. In outline, the combining assembler first examines the set of matching Celera reads to determine if there are excessive pileups indicative of unscreened repetitive elements. Wherever these occur, reads in the repeat region whose mates have not been mapped to consistent positions are removed. Then all sets of mate pairs that consistently imply the same relative position of two bactigs are bundled into a link and weighted according to the number of mates in the bundle. A "greedy" strategy then attempts to order the bactigs by selecting bundles of mate-pairs in order of their weight. A selected mate-pair bundle can tie together two formative scaffolds. It is incorporated to form a single scaffold only if it is consistent with the majority of links between contigs of the scaffold. Once scaffolding is complete, gaps are filled by the "Stones" strategy described above for the WGA assembler.

The GenBank data for the Phase 1 and 2 BACs consisted of an average of 19.8 bactigs per BAC of average size 8099 bp. Application of the combining assembler resulted in individual Celera BAC assemblies being put together into an average of 1.83 scaffolds (median of 1 scaffold) consisting of an average of 8.57 contigs of average size 18,973 bp. In addition to defining order and orientation of the sequence fragments, there were 57% fewer gaps in the combined result. For Phase 0 data, the average GenBank entry consisted of 91.52 reads of average length 784 bp. Application of the combining assembler resulted in an average of 54.8 scaffolds consisting of an average of 58.1 contigs of average size 873 bp. Basically, some small amount of assembly took place, but not enough Celera data were matched to truly assemble the 0.5× to 1× data set represented by the typical Phase 0 BACs. The combining assembler was also applied to the Phase 3 BACs for SNP identification, confirmation of assembly, and localization of the Celera reads. The phase 0 data suggest that a combined whole-genome shotgun data set and 1× light-shotgun of BACs will not yield good assembly of BAC regions; at least 3× light-shotgun of each BAC is needed.

The 5.89 million Celera fragments not matching the GenBank data were assembled with our whole-genome assembler. The assembly resulted in a set of scaffolds totaling 442 Mbp in span and consisting of 326 Mbp of sequence. More than 20% of the scaffolds were >5 kbp long, and these averaged 63% sequence and 27% gaps with a total of 302 Mbp of sequence. All scaffolds >5 kbp were forwarded along with all scaffolds produced by the combining assembler to the subsequent tiling phase.

At this stage, we typically had one or two scaffolds for every BAC region constituting at least 95% of the relevant sequence, and a collection of disjoint Celera-unique scaffolds. The next step in developing the genome components was to determine the order and overlap tiling of these BAC and Celera-unique scaffolds across the genome. For this, we used Celera's 50-kbp mate-pairs information, and BAC-end pairs (*18*) and sequence tagged site (STS) markers (*44*) to provide long-range guidance and chromosome separation. Given the relatively manageable number of scaffolds, we chose not to produce this tiling in a fully automated manner, but to compute an initial tiling with a good heuristic and then use human curators to resolve discrepancies or missed join opportunities. To this end, we developed a graphical user interface that displayed the graph of tiling overlaps and the evidence for each. A human curator could then explore the implication of mapped STS data, dot-plots of sequence overlap, and a visual display of the mate-pair evidence supporting a given choice. The result of this process was a collection of "components," where each component was a tiled set of BAC and Celera-unique scaffolds that had been curator-approved. The process resulted in 3845 components with an estimated span of 2.922 Gbp.

In order to generate the final CSA, we assembled each component with the WGA algorithm. As was done in the WGA process, the bactig data were shredded into a synthetic 2× shotgun data set in order to give the assembler the freedom to independently assemble the data. By using faux reads rather than bactigs, the assembly algorithm could correct errors in the assembly of bactigs and remove chimeric content in a PFP data entry. Chimeric or contaminating sequence (from another part of the genome) would not be incorporated into the reassembly of the component because it did not belong there. In effect, the previous steps in the CSA process served only to bring together Celera fragments and PFP data relevant to a large contiguous segment of the genome, wherein we applied the assembler used for WGA to produce an ab initio assembly of the region.

WGA assembly of the components resulted in a set of scaffolds totaling 2.906 Gbp in span and consisting of 2.654 Gbp of sequence. The chaff, or set of reads not incorporated into the assembly, numbered 6.17 million, or 22%. More than 90.0% of the genome was covered by scaffolds spanning >100 kbp long, and these averaged 92.2% sequence and 7.8% gaps with a total of 2.492 Gbp of sequence. There were a total of 105,264 gaps among the 107,199 contigs that belong to the 1940 scaffolds spanning >100 kbp. The average scaffold size was 1.4 Mbp, the average contig size was 23.24 kbp, and the average gap size was 2.0 kbp where each distribution of sizes was exponential. As such, averages tend to be underrepresentative of the majority of the data. Figure 5 shows a histogram of the bases in scaffolds of various size ranges. Consider also that more than 49% of all gaps were <500 bp long, more than 62% of all gaps were <1 kbp, and all gaps are <100 kbp long. Similarly, more than 73% of the sequence is in contigs > 30 kbp, more than 49% is in contigs >100 kbp, and the largest contig was 1.99 Mbp long. Table 3 provides summary statistics for the structure of this assembly with a direct comparison to the WGA assembly.

2.5 Comparison of the WGA and CSA scaffolds

Having obtained two assemblies of the human genome via independent computational processes (WGA and CSA), we compared scaffolds from the two assemblies as another means of investigating their completeness, consistency, and contiguity. From each assembly, a set of reference scaffolds containing at least 1000 fragments (Celera sequencing reads or bactig shreds) was obtained; this amounted to 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 Gbp and 2.474 Gbp. The sequence of each reference scaffold was compared to the sequence of all scaffolds from the other assembly with which it shared at least 20 fragments or at least 20% of the fragments of the smaller scaffold. For each such comparison, all matches of at least 200 bp with at most 2% mismatch were tabulated.

From this tabulation, we estimated the amount of unique sequence in each assembly in two ways. The first was to determine the number of bases of each assembly that were

not covered by a matching segment in the other assembly. Some 82.5 Mbp of the WGA (3.95%) was not covered by the CSA, whereas 204.5 Mbp (8.26%) of the CSA was not covered by the WGA. This estimate did not require any consistency of the assemblies or any uniqueness of the matching segments. Thus, another analysis was conducted in which matches of less than 1 kbp between a pair of scaffolds were excluded unless they were confirmed by other matches having a consistent order and orientation. This gives some measure of consistent coverage: 1.982 Gbp (95.00%) of the WGA is covered by the CSA, and 2.169 Gbp (87.69%) of the CSA is covered by the WGA by this more stringent measure.

The comparison of WGA to CSA also permitted evaluation of scaffolds for structural inconsistencies. We looked for instances in which a large section of a scaffold from one assembly matched only one scaffold from the other assembly, but failed to match over the full length of the overlap implied by the matching segments. An initial set of candidates was identified automatically, and then each candidate was inspected by hand. From this process, we identified 31 instances in which the assemblies appear to disagree in a nonlocal fashion. These cases are being further evaluated to determine which assembly is in error and why.

In addition, we evaluated local inconsistencies of order or orientation. The following results exclude cases in which one contig in one assembly corresponds to more than one overlapping contig in the other assembly (as long as the order and orientation of the latter agrees with the positions they match in the former). Most of these small rearrangements involved segments on the order of hundreds of base pairs and rarely >1 kbp. We found a total of 295 kbp (0.012%) in the CSA assemblies that were locally inconsistent with the WGA assemblies, whereas 2.108 Mbp (0.11%) in the WGA assembly were inconsistent with the CSA assembly.

The CSA assembly was a few percentage points better in terms of coverage and slightly more consistent than the WGA, because it was in effect performing a few thousand shotgun assemblies of megabase-sized problems, whereas the WGA is performing a shotgun assembly of a gigabase-sized problem. When one considers the increase of two-and-a-half orders of magnitude in problem size, the information loss between the two is remarkably small. Because CSA was logistically easier to deliver and the better of the two results available at the time when downstream analyses needed to be begun, all subsequent analysis was performed on this assembly.

2.6 Mapping scaffolds to the genome

The final step in assembling the genome was to order and orient the scaffolds on the chromosomes. We first grouped scaffolds together on the basis of their order in the components from CSA. These grouped scaffolds were reordered by examining residual mate-pairing data between the scaffolds. We next mapped the scaffold groups onto the chromosome using physical mapping data. This step depends on having reliable high-resolution map information such that each scaffold will overlap multiple markers. There are two genome-wide types of map information available: high-density STS maps and fingerprint maps of BAC clones developed at Washington University (*45*). Among the genome-wide STS maps, GeneMap99 (GM99) has the most markers and therefore was most useful for mapping scaffolds. The two different mapping approaches are complementary to one another. The fingerprint maps should have better local order because they were built by comparison of overlapping BAC clones. On the other hand, GM99 should have a more reliable long-range order, because the framework markers were derived from well-validated genetic maps. Both types of maps were used as a reference for human curation of the components that were the input to the regional assembly, but they did not determine the order of sequences produced by the assembler.

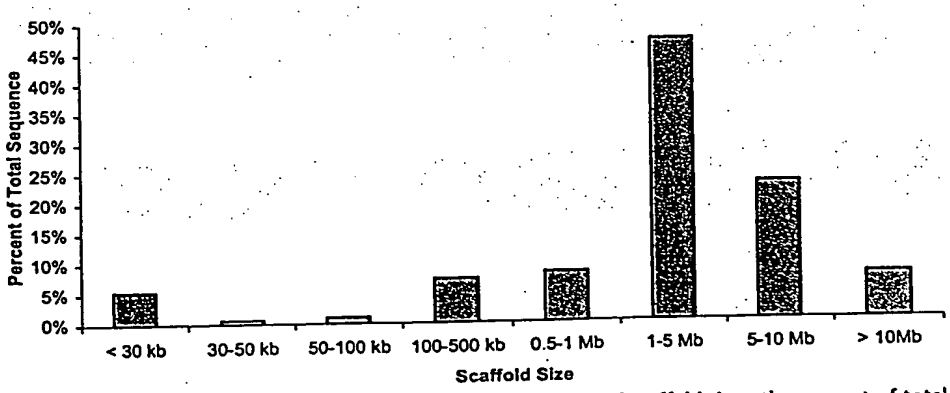


Fig. 5. Distribution of scaffold sizes of the CSA. For each range of scaffold sizes, the percent of total sequence is indicated.

In order to determine the effectiveness of the fingerprint maps and GM99 for mapping scaffolds, we first examined the reliability of these maps by comparison with large scaffolds. Only 1% of the STS markers on the 10 largest scaffolds (those >9 Mbp) were mapped on a different chromosome on GM99. Two percent of the STS markers disagreed in position by more than five framework bins. However, for the fingerprint maps, a 2% chromosome discrepancy was observed, and on average 23.8% of BAC locations in the scaffold sequence disagreed with fingerprint map placement by more than five BACs. When further examining the source of discrepancy, it was found that most of the discrepancy came from 4 of the 10 scaffolds, indicating this there is variation in the quality of either the map or the scaffolds. All four scaffolds were assembled, as well as the other six, as judged by clone coverage analysis, and showed the same low discrepancy rate to GM99, and thus we concluded that the fingerprint map global order in these cases was not reliable. Smaller scaffolds had a higher discordance rate with GM99 (4.21% of STSs were discordant by more than five framework bins), but a lower discordance rate with the fingerprint maps (11% of BACs disagreed with fingerprint maps by more than five BACs). This observation agrees with the clone coverage analysis (*46*) that Celera scaffold construction was better supported by long-range mate pairs in larger scaffolds than in small scaffolds.

We created two orderings of Celera scaffolds on the basis of the markers (BAC or STS) on these maps. Where the order of scaffolds agreed between GM99 and the WashU BAC map, we had a high degree of confidence that that order was correct; these scaffolds were termed "anchor scaffolds." Only scaffolds with a low overall discrepancy rate with both maps were considered anchor scaffolds. Scaffolds in GM99 bins were allowed to permute in their order to match WashU ordering, provided they did not violate their framework orders. Orientation of individual scaffolds was determined by the presence of multiple mapped markers with consistent order. Scaffolds with only one marker have insufficient information to assign orientation. We found 70.1% of the genome in anchored scaffolds, more than 99% of which are also oriented (Table 4). Because GM99 is of lower resolution than the WashU map, a number of scaffolds without STS matches could be ordered relative to the anchored scaffolds because they included sequence from the same or adjacent BACs on the WashU map. On the other hand, because of occasional WashU global ordering discrepancies, a number of scaffolds determined to be "unmappable" on the WashU map could be ordered relative to the anchored scaffolds

with GM99. These scaffolds were termed "ordered scaffolds." We found that 13.9% of the assembly could be ordered by these additional methods, and thus 84.0% of the genome was ordered unambiguously.

Next, all scaffolds that could be placed, but not ordered, between anchors were assigned to the interval between the anchored scaffolds and were deemed to be "bounded" between them. For example, small scaffolds having STS hits from the same GeneMap bin or hitting the same BAC cannot be ordered relative to each other, but can be assigned a placement boundary relative to other anchored or ordered scaffolds. The remaining scaffolds either had no localization information, conflicting information, or could only be assigned to a generic chromosome location. Using the above approaches, ~98% of the genome was anchored, ordered, or bounded.

Finally, we assigned a location for each scaffold placed on the chromosome by spreading out the scaffolds per chromosome. We assumed that the remaining unmapped scaffolds, constituting 2% of the genome, were distributed evenly across the genome. By dividing the sum of unmapped scaffold lengths with the sum of the number of mapped scaffolds, we arrived at an estimate of interscaffold gap of 1483 bp. This gap was used to separate all the scaffolds on each chromosome and to assign an offset in the chromosome.

During the scaffold-mapping effort, we encountered many problems that resulted in additional quality assessment and validation analysis. At least 978 (3% of 33,173) BACs were believed to have sequence data from more than one location in the genome (*47*). This is consistent with the bactig chimerism analysis reported above in the Assembly Strategies section. These BACs could not be assigned to unique positions within the CSA assembly and thus could not be used for ordering scaffolds. Likewise, it was not always possible to assign STSs to unique locations in the assembly because of genome duplications, repetitive elements, and pseudogenes.

Because of the time required for an exhaustive search for a perfect overlap, CSA generated 21,607 intrascaffold gaps where the mate-pair data suggested that the contigs should overlap, but no overlap was found. These gaps were defined as a fixed 50 bp in length and make up 18.6% of the total 116,442 gaps in the CSA assembly.

We chose not to use the order of exons implied in cDNA or EST data as a way of ordering scaffolds. The rationale for not using this data was that doing so would have biased certain regions of the assembly by rearranging scaffolds to fit the transcript data and made validation of both the assembly and gene definition processes more difficult.

2.7 Assembly and validation analysis

We analyzed the assembly of the genome from the perspectives of completeness (amount of coverage of the genome) and correctness (the structural accuracy of the order and orientation and the consensus sequence of the assembly).

Completeness. Completeness is defined as the percentage of the euchromatic sequence represented in the assembly. This cannot be known with absolute certainty until the euchromatin sequence has been completed. However, it is possible to estimate completeness on the basis of (i) the estimated sizes of intrascaffold gaps; (ii) coverage of the two published chromosomes, 21 and 22 (*48*, *49*); and (iii) analysis of the percentage of an independent set of random sequences (STS markers) contained in the assembly. The whole-genome libraries contain heterochromatic sequence and, although no attempt has been made to assemble it, there may be instances of unique sequence embedded in regions of heterochromatin as were observed in *Drosophila* (*50*, *51*).

The sequences of human chromosomes 21 and 22 have been completed to high quality and published (*48*, *49*). Although this sequence served as input to the assembler, the finished sequence was shredded into a shotgun data set so that the assembler had the opportunity to assemble it differently from the original sequence in the case of structural polymorphisms or assembly errors in the BAC data. In particular, the assembler must be able to resolve repetitive elements at the scale of components (generally multimegabase in size), and so this comparison reveals the level to which the assembler resolves repeats. In certain areas, the assembly structure differs from the published versions of chromosomes 21 and 22 (see below). The consequence of the flexibility to assemble "finished" sequence differently on the basis of Celera data resulted in an assembly with more segments than the chromosome 21 and 22 sequences. We examined the reasons why there are more gaps in the Celera sequence than in chromosomes 21 and 22 and expect that they may be typical of gaps in other regions of the genome. In the Celera assembly, there are 25 scaffolds, each containing at least 10 kb of sequence, that collectively span 94.3% of chromosome 21. Sixty-two scaffolds span 95.7% of chromosome 22. The total length of the gaps remaining in the Celera assembly for these two chromosomes is 3.4 Mbp. These gap sequences were analyzed by RepeatMasker and by searching against the entire genome assembly (*52*). About 50% of the gap sequence consisted of common repetitive elements identified by RepeatMasker; more than half of the remainder was lower copy number repeat elements.

A more global way of assessing completeness is to measure the content of an independent set of sequence data in the assembly. We compared 48,938 STS markers from Genemap99 (*51*) to the scaffolds. Because these markers were not used in the assembly processes, they provided a truly independent measure of completeness. ePCR (*53*) and BLAST (*54*) were used to locate STSs on the assembled genome. We found 44,524 (91%) of the STSs in the mapped genome. An additional 2648 markers (5.4%) were found by searching the unassembled data or "chaff." We identified 1283 STS markers (2.6%) not found in either Celera sequence or BAC data as of September 2000, raising the possibility that these markers may not be of human origin. If that were the case, the Celera assembled sequence would represent 93.4% of the human genome and the unassembled data 5.5%, for a total of 98.9% coverage. Similarly, we compared CSA against 36,678 TNG radiation hybrid markers (*55a*) using the same method. We found that 32,371 markers (88%) were located in the mapped CSA scaffolds, with 2055 markers (5.6%) found in the remainder. This gave a 94% coverage of the genome through another genome-wide survey.

Correctness. Correctness is defined as the structural and sequence accuracy of the assembly. Because the source sequences for the Celera data and the GenBank data are from different individuals, we could not directly compare the consensus sequence of the as-

Table 4. Summary of scaffold mapping. Scaffolds were mapped to the genome with different levels of confidence (anchored scaffolds have the highest confidence; unmapped scaffolds have the lowest). Anchored scaffolds were consistently ordered by the WashU BAC map and GM99. Ordered scaffolds were consistently ordered by at least one of the following: the WashU BAC map, GM99, or component tiling path. Bounded scaffolds had order conflicts between at least two of the external maps, but their placements were adjacent to a neighboring anchored or ordered scaffold. Unmapped scaffolds had, at most, a chromosome assignment. The scaffold subcategories are given below each category.

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

sembly against other finished sequence for determining sequencing accuracy at the nucleotide level, although this has been done for identifying polymorphisms as described in Section 6. The accuracy of the consensus sequence is at least 99.96% on the basis of a statistical estimate derived from the quality values of the underlying reads.

The structural consistency of the assembly can be measured by mate-pair analysis. In a correct assembly, every mated pair of sequencing reads should be located on the consensus sequence with the correct separation and orientation between the pairs. A pair is termed "valid" when the reads are in the correct orientation and the distance between them is within the mean ± 3 standard deviations of the distribution of insert sizes of the library from which the pair was sampled. A pair is termed "misoriented" when the reads are not correctly oriented, and is termed "misseparated" when the distance between the reads is not in the correct range but the reads are correctly oriented. The mean ± the standard deviation of each library used by the assembler was determined as described above. To validate these, we examined all reads mapped to the finished sequence of chromosome 21 (*48*) and determined how many incorrect mate pairs there were as a result of laboratory tracking errors and chimerism (two different segments of the genome cloned into the same plasmid), and how tight the distribution of insert sizes was for those that were correct (Table 5). The standard deviations for all Celera libraries were quite small, less than 15% of the insert length, with the exception of a few 50-kbp libraries. The 2- and 10-kbp libraries contained less than 2% invalid mate pairs, whereas the 50-kbp libraries were somewhat higher (~10%). Thus, although the mate-pair information was not perfect, its accuracy was such that measuring valid, misoriented, and misseparated pairs with respect to a given assembly was deemed to be a reliable instrument for validation purposes, especially when several mate pairs confirm or deny an ordering.

The clone coverage of the genome was 39×, meaning that any given base pair was, on average, contained in 39 clones or, equivalently, spanned by 39 mate-paired reads. Areas of low clone coverage or areas with a high proportion of invalid mate pairs would indicate potential assembly problems. We computed the coverage of each base in the assembly by valid mate pairs (Table 6). In summary, for scaffolds >30 kbp in length, less than 1% of the Celera assembly was in regions of less than 3× clone coverage. Thus, more than 99% of the assembly, including order and orientation, is strongly supported by this measure alone.

We examined the locations and number of all misoriented and misseparated mates. In addition to doing this analysis on the CSA assembly (as of 1 October 2000), we also performed a study of the PFP assembly as of 5 September 2000 (*30*, *55b*). In this latter case, Celera mate pairs had to be mapped to the PFP assembly. To avoid mapping errors due to high-fidelity repeats, the only pairs mapped were those for which both reads matched at only one location with less than 6% differences. A threshold was set such that sets of five or more simultaneously invalid mate pairs indicated a potential breakpoint, where the construction of the two assemblies differed. The graphic comparison of the CSA chromosome 21 assembly with the published sequence (Fig. 6A) serves as a validation of this methodology. Blue tick marks in the panels indicate breakpoints. There were a similar (small) number of breakpoints on both chromosome sequences. The exception was 12 sets of scaffolds in the Celera assembly (a total of 3% of the chromosome length in 212 single-contig scaffolds) that were mapped to the wrong positions because they were too small to be mapped reliably. Figures 6 and 7 and Table 6 illustrate the mate-pair differences and breakpoints between the two assemblies. There was a higher percentage of misoriented and misseparated mate pairs in the large-insert libraries (50 kbp and BAC ends) than in the small-insert libraries in both assemblies (Table 6). The large-insert libraries are more likely to identify discrepancies simply because they span a larger segment of the genome. The graphic comparison between the two assemblies for chromosome 8 (Fig. 6, B and C) shows that there are many

Table 5. Mate-pair validation. Celera fragment sequences were mapped to the published sequence of chromosome 21. Each mate pair uniquely mapped was evaluated for correct orientation and placement (number of mate pairs tested). If the two mates had incorrect relative orientation or placement, they were considered invalid (number of invalid mate pairs).

		Chromosome 21						Genome		
Library type	Library no.	Mean insert size (bp)	SD (bp)	SD/ mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/ mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768 (mean = 2.7)	2.7			

more breakpoints for the PFP assembly than for the Celera assembly. Figure 7 shows the breakpoint map (blue tick marks) for both assemblies of each chromosome in a side-by-side fashion. The order and orientation of Celera's assembly shows substantially fewer breakpoints except on the two finished chromosomes. Figure 7 also depicts large gaps (>10 kbp) in both assemblies as red tick marks. In the CSA assembly, the size of all gaps have been estimated on the basis of the mate-pair data. Breakpoints can be caused by structural polymorphisms, because the two assemblies were derived from different human genomes. They also reflect the unfinished nature of both genome assemblies.

3 Gene Prediction and Annotation

Summary. To enumerate the gene inventory, we developed an integrated, evidence-based approach named Otto. The evidence used to increase the likelihood of identifying genes includes regions conserved between the mouse and human genomes, similarity to ESTs or other mRNA-derived data, or similarity to other proteins. A comparison of Otto (combined Otto-RefSeq and Otto homology) with Genscan, a standard gene-prediction algorithm, showed greater sensitivity (0.78 versus 0.50) and specificity (0.93 versus 0.63) of Otto in the ability to define gene structure. Otto-predicted genes were complemented with a set of genes from three gene-prediction programs that exhibited weaker, but still significant, evidence that they may be expressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections. Extensive manual curation to establish precise characterization of gene structure will be necessary to improve the results from this initial computational approach.

3.1 Automated gene annotation

A gene is a locus of cotranscribed exons. A single gene may give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by means of alternative splicing and alternative transcription initiation and termination sites. Our cells are able to discern within the billions of base pairs of the genomic DNA the signals for initiating transcription and for splicing together exons separated by a few or hundreds of thousands of base pairs. The first step in characterizing the genome is to define the structure of each gene and each transcription unit.

The number of protein-coding genes in mammals has been controversial from the outset. Initial estimates based on reassociation data placed it between 30,000 to 40,000, whereas later estimates from the brain were >100,000 (*56*). More recent data from both the corporate and public sectors, based on extrapolations from EST, CpG island, and transcript density–based extrapolations, have not reduced this variance. The highest recent number of 142,634 genes emanates from a report from Incyte Pharmaceuticals, and is based on a combination of EST data and the association of ESTs with CpG islands (*57*). In stark contrast are three quite different, and much lower estimates: one of ~35,000 genes derived with genome-wide EST data and sampling procedures in conjunction with chromosome 22 data (*58*); another of 28,000 to 34,000 genes derived with a comparative methodology involving sequence conservation between humans and the puffer fish *Tetraodon nigroviridis* (*59*); and a figure of 35,000 genes, which was derived simply by extrapolating from the density of 770 known and predicted genes in the 67 Mbp of chromosomes 21 and 22, to the approximately 3-Gbp euchromatic genome.

The problem of computational identification of transcriptional units in genomic DNA sequence can be divided into two phases. The first is to partition the sequence into segments that are likely to correspond to individual genes. This is not trivial and is a weakness of most de novo gene-finding algorithms. It is also critical to determining the number of genes in the human gene inventory. The second challenge is to construct a gene model that reflects the probable structure of the transcript(s) encoded in the region. This can be done with reasonable accuracy when a full-length cDNA has been sequenced or a highly homologous protein sequence is known. De novo gene prediction, although less accurate, is the only way to find genes that are not represented by homologous proteins or ESTs. The following section describes the methods we have developed to address these problems for the prediction of protein-coding genes.

We have developed a rule-based expert system, called Otto, to identify and characterize genes in the human genome (*60*). Otto attempts to simulate in software the process that a human annotator uses to identify a gene and refine its structure. In the process of annotating a region of the genome, a human curator examines the evidence provided by the computational pipeline (described below) and examines how various types of evidence relate to one another. A curator puts different levels of confidence in different types of evidence and looks for certain patterns of evidence to support gene annotation. For example, a curator may examine homology to a number of ESTs and evaluate whether or not they can be connected into a longer, virtual mRNA. The curator would also evaluate the strength of the similarity and the contiguity of the match, in essence asking whether any ESTs cross splice-junctions and whether the edges of putative exons have consensus splice sites. This kind of manual annotation process was used to annotate the *Drosophila* genome.

The Otto system can promote observed evidence to a gene annotation in one of two ways. First, if the evidence includes a high-quality match to the sequence of a known gene [here defined as a human gene represented in a curated subset of the RefSeq database (*61*)], then Otto can promote this to a gene annotation. In the second method, Otto evaluates a broad spectrum of evidence and determines if this evidence is adequate to support promotion to a gene annotation. These processes are described below.

Initially, gene boundaries are predicted on the basis of examination of sets of overlapping protein and EST matches generated by a computational pipeline (*62*). This pipeline searches the scaffold sequences against protein, EST, and genome-sequence databases to define regions of sequence similarity and runs three de novo gene-prediction programs.

To identify likely gene boundaries, regions of the genome were partitioned by Otto on the basis of sequence matches identified by BLAST. Each of the database sequences matched in the region under analysis was compared by an algorithm that takes into account both coordinates of the matching sequence, as well as the sequence type (e.g. protein, EST, and so forth). The results were used to group the matches into bins of relate sequences that may define a gene and identif

Table 6. Genome-wide mate pair analysis of compartmentalized shotgun (CSA) and PFP assemblies.*

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated†	% valid	% mis-oriented	% mis-separated†
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

*Data for individual chromosomes can be found in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1. †Mates are misseparated if their distance is >3 SD from the mean library size.

gene boundaries. During this process, multiple hits to the same region were collapsed to a coherent set of data by tracking the coverage of a region. For example, if a group of bases was represented by multiple overlapping ESTs, the union of these regions matched by the set of ESTs on the scaffold was marked as being supported by EST evidence. This resulted in a series of "gene bins," each of which was believed to contain a single gene. One weakness of this initial implementation of the algorithm was in predicting gene boundaries in regions of tandemly duplicated genes. Gene clusters frequently resulted in homologous neighboring genes being joined together, resulting in an annotation that artificially concatenated these gene models.

Next, known genes (those with exact matches of a full-length cDNA sequence to the genome) were identified, and the region corresponding to the cDNA was annotated as a predicted transcript. A subset of the curated human gene set RefSeq from the National Center for Biotechnology Information (NCBI) was included as a data set searched in the computational pipeline. If a RefSeq transcript matched the genome assembly for at least 50% of its length at >92% identity, then the SIM4 (*63*) alignment of the RefSeq transcript to the region of the genome under analysis was promoted to the status of an Otto annotation. Because the genome sequence has gaps and sequence errors such as frameshifts, it was not always possible to predict a transcript that agrees precisely with the experimentally determined cDNA sequence. A total of 6538 genes in our inventory were identified and transcripts predicted in this way.

Regions that have a substantial amount of sequence similarity, but do not match known genes, were analyzed by that part of the Otto system that uses the sequence similarity information to predict a transcript. Here, Otto

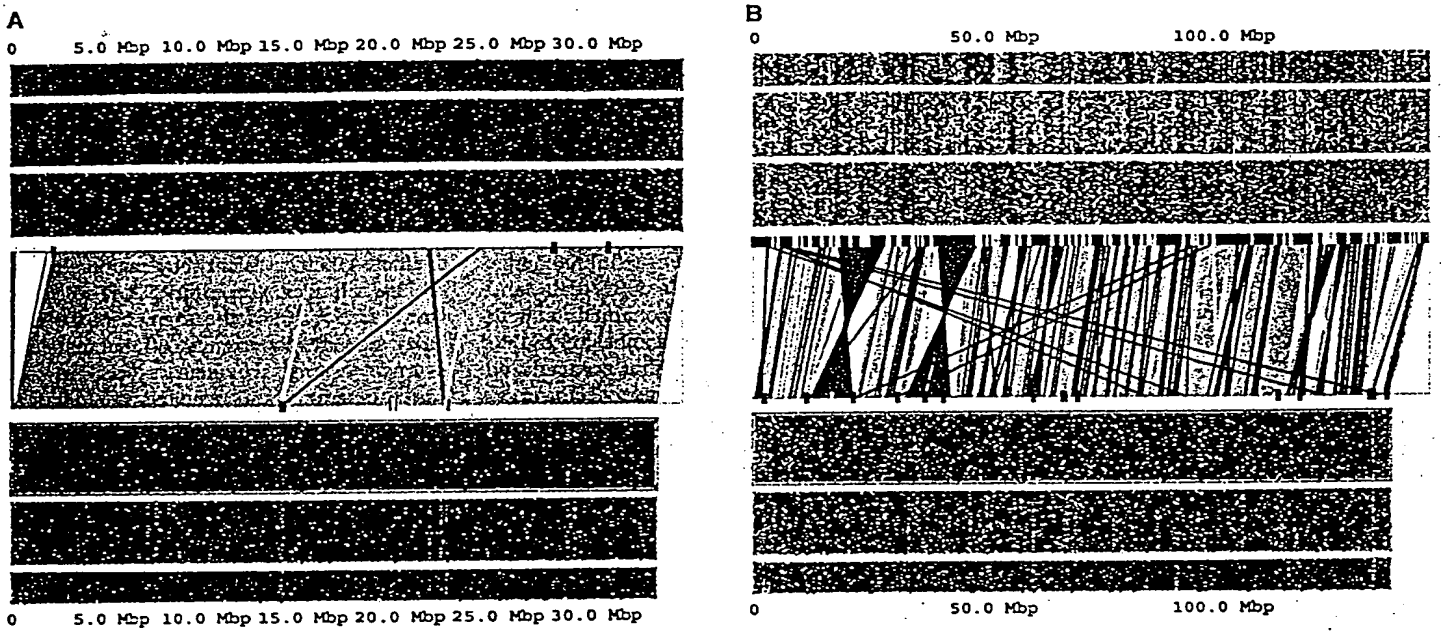


Fig. 6. Comparison of the CSA and the PFP assembly. (A) All of chromosome 21, (B) all of chromosome 8, and (C) a 1-Mb region of chromosome 8 representing a single Celera scaffold. To generate the figure, Celera fragment sequences were mapped onto each assembly. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50 kbp long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars. Plots of all 24 chromosomes can be seen in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1.

evaluates evidence generated by the computational pipeline, corresponding to conservation between mouse and human genomic DNA, similarity to human transcripts (ESTs and cDNAs), similarity to rodent transcripts (ESTs and cDNAs), and similarity of the translation of human genomic DNA to known proteins to predict potential genes in the human genome. The sequence from the region of genomic DNA contained in a gene bin was extracted, and the subsequences supported by any homology evidence were marked (plus 100

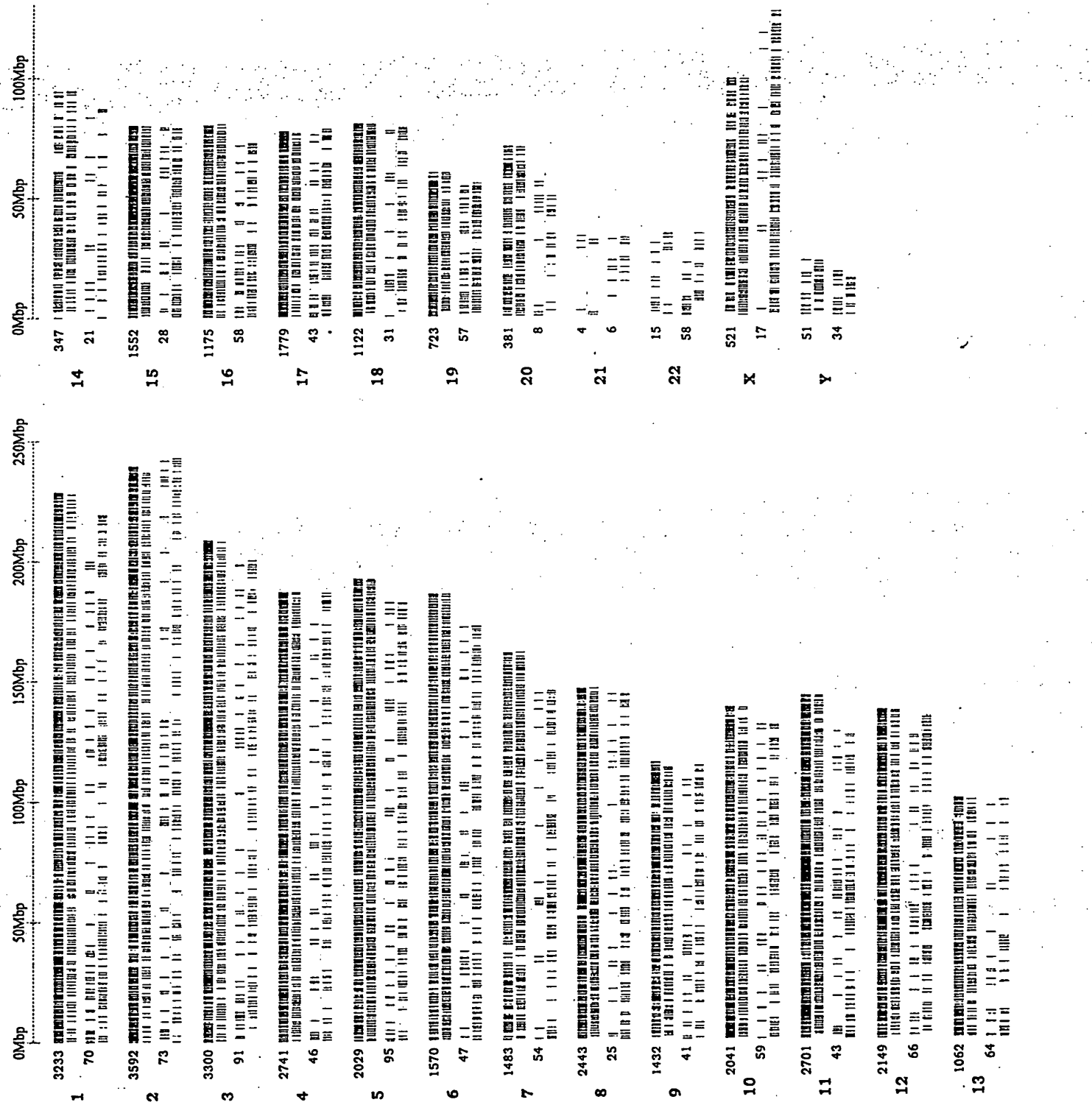


Fig. 7. Schematic view of the distribution of breakpoints and large gaps on all chromosomes. For each chromosome, the upper pair of lines represent the PFP assembly, and the lower pair of lines represent Celera's assembly. Blue tick marks represent breakpoints, whereas red tick marks represent a gap of larger than 10,000 bp. The number of breakpoints per chromosome is indicated in black, and the chromosome numbers in red.

bases flanking these regions). The other bases in the region, those not covered by any homology evidence, were replaced by N's. This sequence segment, with high confidence regions represented by the consensus genomic sequence and the remainder represented by N's, was then evaluated by Genscan to see if a consistent gene model could be generated. This procedure simplified the gene-prediction task by first establishing the boundary for the gene (not a strength of most gene-finding algorithms), and by eliminating regions with no supporting evidence. If Genscan returned a plausible gene model, it was further evaluated before being promoted to an "Otto" annotation. The final Genscan predictions were often quite different from the prediction that Genscan returned on the same region of native genomic sequence. A weakness of using Genscan to refine the gene model is the loss of valid, small exons from the final annotation.

The next step in defining gene structures based on sequence similarity was to compare each predicted transcript with the homology-based evidence that was used in previous steps to evaluate the depth of evidence for each exon in the prediction. Internal exons were considered to be supported if they were covered by homology evidence to within ±10 bases of their edges. For first and last exons, the internal edge was required to be within 10 bases, but the external edge was allowed greater latitude to allow for 5′ and 3′ untranslated regions (UTRs). To be retained, a prediction for a multi-exon gene must have evidence such that the total number of "hits," as defined above, divided by the number of exons in the prediction must be >0.66 or must correspond to a RefSeq sequence. A single-exon gene must be covered by at least three supporting hits (±10 bases on each side), and these must cover the complete predicted open reading frame. For a single-exon gene, we also required that the Genscan prediction include both a start and a stop codon. Gene models that did not meet these criteria were disregarded, and those that passed were promoted to Otto predictions. Homology-based Otto predictions do not contain 3′ and 5′ untranslated sequence. Although three de novo gene-finding programs [GRAIL, Genscan, and FgenesH (*63*)] were run as part of the computational analysis, the results of these programs were not directly used in making the Otto predictions. Otto predicted 11,226 additional genes by means of sequence similarity.

Table 7. Sensitivity and specificity of Otto and Genscan. Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; $P < 0.001$).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

*Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. †Refers to those annotations produced by supplying all available evidence to Genscan.

3.2 Otto validation

To validate the Otto homology-based process and the method that Otto uses to define the structures of known genes, we compared transcripts predicted by Otto with their corresponding (and presumably correct) transcript from a set of 4512 RefSeq transcripts for which there was a unique SIM4 alignment (Table 7). In order to evaluate the relative performance of Otto and Genscan, we made three comparisons. The first involved a determination of the accuracy of gene models predicted by Otto with only homology data other than the corresponding RefSeq sequence (Otto homology in Table 7). We measured the sensitivity (correctly predicted bases divided by the total length of the cDNA) and specificity (correctly predicted bases divided by the sum of the correctly and incorrectly predicted bases). Second, we examined the sensitivity and specificity of the Otto predictions that were made solely with the RefSeq sequence, which is the process that Otto uses to annotate known genes (Otto-RefSeq). And third, we determined the accuracy of the Genscan predictions corresponding to these RefSeq sequences. As expected, the alignment method (Otto-RefSeq) was the most accurate, and Otto-homology performed better than Genscan by both criteria. Thus, 6.1% of true RefSeq nucleotides were not represented in the Otto-refseq annotations and 2.7% of the nucleotides in the Otto-RefSeq transcripts were not contained in the original RefSeq transcripts. The discrepancies could come from legitimate differences between the Celera assembly and the RefSeq transcript due to polymorphisms, incomplete or incorrect data in the Celera assembly, errors introduced by Sim4 during the alignment process, or the presence of alternatively spliced forms in the data set used for the comparisons.

Because Otto uses an evidence-based approach to reconstruct genes, the absence of experimental evidence for intervening exons may inadvertantly result in a set of exons that cannot be spliced together to give rise to a transcript. In such cases, Otto may "split genes" when in fact all the evidence should be combined into a single transcript. We also examined the tendency of these methods to incorrectly split gene predictions. These trends are shown in Fig. 8. Both RefSeq and homology-based predictions by Otto split known genes into fewer segments than Genscan alone.

3.3 Gene number

Recognizing that the Otto system is quite conservative, we used a different gene-prediction strategy in regions where the homology evidence was less strong. Here the results of de novo gene predictions were used. For these genes, we insisted that a predicted transcript have at least two of the following types of evidence to be included in the gene set for further analysis: protein, human EST, rodent EST, or mouse genome fragment matches. This final class of predicted genes is a subset of the predictions made by the three gene-finding programs that were used in the computational pipeline. For these, there was not sufficient sequence similarity information for Otto to attempt to predict a gene structure. The three de novo gene-finding programs resulted in about 155,695 predictions, of which ~76,410 were nonredundant (nonoverlapping with one another). Of these, 57,935 did not overlap known genes or predictions made by Otto. Only 21,350 of the gene predictions that did not overlap Otto predictions were partially supported by at least one type of sequence similarity evidence, and 8619 were partially supported by two types of evidence (Table 8).

The sum of this number (21,350) and the number of Otto annotations (17,764), 39,114, is near the upper limit for the human gene complement. As seen in Table 8, if the requirement for other supporting evidence is made more stringent, this number drops rapidly so that demanding two types of evidence reduces the total gene number to 26,383 and demanding three types reduces it to ~23,000. Requiring that a prediction be supported by all four categories of evidence is too stringent because it would eliminate genes that encode novel proteins (members of currently undescribed protein families). No correction for pseudogenes has been made at this point in the analysis.

In a further attempt to identify genes that were not found by the autoannotation process or any of the de novo gene finders, we examined regions outside of gene predictions that were similar to the EST sequence, and where the EST matched the genomic sequence across a splice junction. After correcting for potential 3′ UTRs of predicted genes, about 2500 such regions remained. Addition of a requirement for at least one of the following evidence types—homology to mouse genomic sequence fragments, rodent ESTs, or cDNAs—or similarity to a known protein reduced this number to 1010. Adding this to the numbers from the previous paragraph would give us estimates of about 40,000, 27,000, and 24,000 potential genes in the human genome, depending on the stringency of evidence considered. Table 8 illustrates the number of genes and presents the degree of

confidence based on the supporting evidence. Transcripts encoded by a set of 26,383 genes were assembled for further analysis. This set includes the 6538 genes predicted by Otto on the basis of matches to known genes, 11,226 transcripts predicted by Otto based on homology evidence, and 8619 from the subset of transcripts from de novo gene-prediction programs that have two types of supporting evidence. The 26,383 genes are illustrated along chromosome diagrams in Fig. 1. These are a very preliminary set of annotations and are subject to all the limitations of an automated process. Considerable refinement is still necessary to improve the accuracy of these transcript predictions. All the predictions and descriptions of genes and the associated evidence that we present are the product of completely computational processes, not expert curation. We have attempted to enumerate the genes in the human genome in such a way that we have different levels of confidence based on the amount of supporting evidence: known genes, genes with good protein or EST homology evidence, and de novo gene predictions confirmed by modest homology evidence.

3.4 Features of human gene transcripts

We estimate the average span for a "typical" gene in the human DNA sequence to be about 27,894 bases. This is based on the average span covered by RefSeq transcripts, used because it represents our highest confidence set.

The set of transcripts promoted to gene annotations varies in a number of ways. As can be seen from Table 8 and Fig. 9, transcripts predicted by Otto tend to be longer, having on average about 7.8 exons, whereas those promoted from gene-prediction programs average about 3.7 exons. The largest number of exons that we have identified in a transcript is 234 in the titin mRNA. Table 8 compares the amounts of evidence that support the Otto and other predicted transcripts. For example, one can see that a typical Otto transcript has 6.99 of its 7.81 exons supported by protein homology evidence. As would be expected, the Otto transcripts generally have more support than do transcripts predicted by the de novo methods.

4 Genome Structure

Summary. This section describes several of the noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set. These include an analysis of G+C content and gene density in the context of cytogenetic maps of the genome, an enumerative analysis of CpG islands, and a brief description of the genome-wide repetitive elements.

4.1 Cytogenetic maps

Perhaps the most obvious, and certainly the most visible, element of the structure of the genome is the banding pattern produced by Giemsa stain. Chromosomal banding studies have revealed that about 17% to 20% of the human chromosome complement consists of C-bands, or constitutive heterochromatin (*64*). Much of this heterochromatin is highly polymorphic and consists of different families of alpha satellite DNAs with various higher order repeat structures (*65*). Many chromosomes have complex inter- and intrachromosomal duplications present in pericentromeric regions (*66*). About 5% of the sequence reads were identified as alpha satellite sequences; these were not included in the assembly.

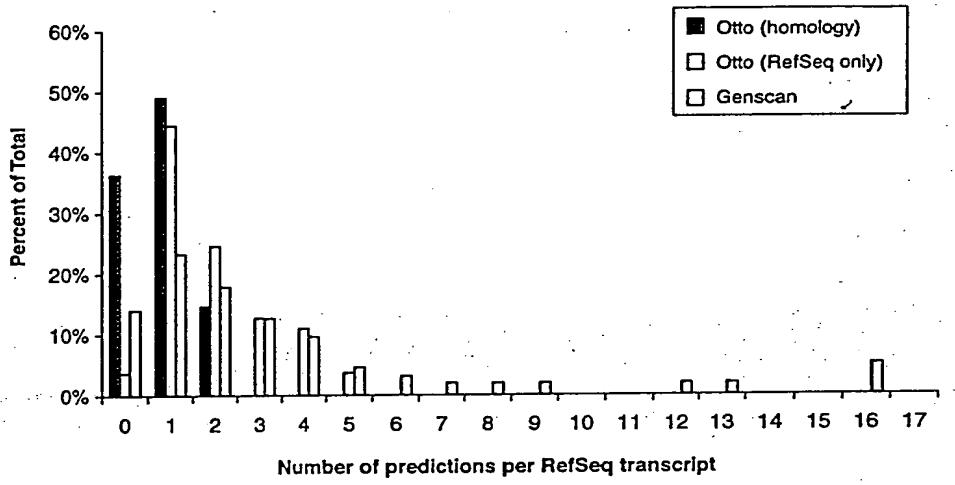


Fig. 8. Analysis of split genes resulting from different annotation methods. A set of 4512 Sim4-based alignments of RefSeq transcripts to the genomic assembly were chosen (see the text for criteria), and the numbers of overlapping Genscan, Otto (RefSeq only) annotations based solely on Sim4-polished RefSeq alignments, and Otto (homology) annotations (annotations produced by supplying all available evidence to Genscan) were tallied. These data show the degree to which multiple Genscan predictions and/or Otto annotations were associated with a single RefSeq transcript. The zero class for the Otto-homology predictions shown here indicates that the Otto-homology calls were made without recourse to the RefSeq transcript, and thus no Otto call was made because of insufficient evidence.

Table 8. Numbers of exons and transcripts supported by various types of evidence for Otto and de novo gene prediction methods. Highlighted cells indicate the gene sets analyzed in this paper (boldface, set of genes selected for protein analysis; italic, total set of accepted de novo predictions).

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	**17,968†**	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	*21,350*	8,619	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per transcript	Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00	4.28
	De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42	3.16

*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

Examination of pericentromeric regions is ongoing.

The remaining ~80% of the genome, the euchromatic component, is divisible into G-, R-, and T-bands (*67*). These cytogenetic bands have been presumed to differ in their nucleotide composition and gene density, although we have been unable to determine precise band boundaries at the molecular level. T-bands are the most G+C- and gene-rich, and G-bands are G+C-poor (*68*). Bernardi has also offered a description of the euchromatin at the molecular level as long stretches of DNA of differing base composition, termed isochores (denoted L, H1, H2, and H3), which are >300 kbp in length (*69*). Bernardi defined the L (light) isochores as G+C-poor (<43%), whereas the H (heavy) isochores fall into three G+C-rich classes representing 24, 8, and 5% of the genome. Gene concentration has been claimed to be very low in the L isochores and 20-fold more enriched in the H2 and H3 isochores (*70*). By examining contiguous 50-kbp windows of G+C content across the assembly, we found that regions of G+C content >48% (H3 isochores) averaged 273.9 kbp in length, those with G+C content between 43 and 48% (H1+H2 isochores) averaged 202.8 kbp in length, and the average span of regions with <43% (L isochores) was 1078.6 kbp. The correlation between G+C content and gene density was also examined in 50-kbp windows along the assembled sequence (Table 9 and Figs. 10 and 11). We found that the density of genes was greater in regions of high G+C than in regions of low G+C content, as expected. However, the correlation between G+C content and gene density was not as skewed as previously predicted (*69*). A higher proportion of genes were located in the G+C-poor regions than had been expected.

Chromosomes 17, 19, and 22, which have a disproportionate number of H3-containing bands, had the highest gene density (Table 10). Conversely, of the chromosomes that we found to have the lowest gene density, X, 4, 18, 13, and Y, also have the fewest H3 bands. Chromosome 15, which also has few H3 bands, did not have a particularly low gene density in our analysis. In addition, chromosome 8, which we found to have a low gene density, does not appear to be unusual in its H3 banding.

How valid is Ohno's postulate (*71*) that mammalian genomes consist of oases of genes in otherwise essentially empty deserts? It appears that the human genome does indeed contain deserts, or large, gene-poor regions. If we define a desert as a region >500 kbp without a gene, then we see that 605 Mbp, or about 20% of the genome, is in deserts. These are not uniformly distributed over the various chromosomes. Gene-rich chromosomes 17, 19, and 22 have only about 12% of their collective 171 Mbp in deserts, whereas gene-poor chromosomes 4, 13, 18, and X have 27.5% of their 492 Mbp in deserts (Table 11). The apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function.

4.2 Linkage map

Linkage maps provide the basis for genetic analysis and are widely used in the study of the inheritance of traits and in the positional cloning of genes. The distance metric, centimorgans (cM), is based on the recombination rate between homologous chromosomes during meiosis. In general, the rate of recombination in females is greater than that in males, and this degree of map expansion is not uniform across the genome (*72*). One of the opportunities enabled by a nearly complete genome sequence is to produce the ultimate physical map, and to fully analyze its correspondence with two other maps that have been widely used in genome and genetic analysis: the linkage map and the cytogenetic map. This would close the loop between the mapping and sequencing phases of the genome project.

We mapped the location of the markers that constitute the Genethon linkage map to the genome. The rate of recombination, expressed as cM per Mbp, was calculated for 3-Mbp windows as shown in Table 12. Higher rates of recombination in the telomeric region of the chromosomes have been previously documented (*73*). From this mapping result, there is a difference of 4.99 between lowest rates and highest rates and the largest difference of 4.4 between males and females (4.99 to 0.47 on chromosome 16). This indicates that the variability in recombination rates among regions of the genome exceeds the differences in recombination rates between males and females. The human genome has recombination hotspots, where recombination rates vary fivefold or more over a space of 1 kbp, so the picture one gets of the magnitude of variability in recombination rate will depend on the size of the window

Table 9. Characteristics of G+C in isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43–48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

*The predictions were based on Bernardi's definitions (*70*) of the isochore structure of the human genome.

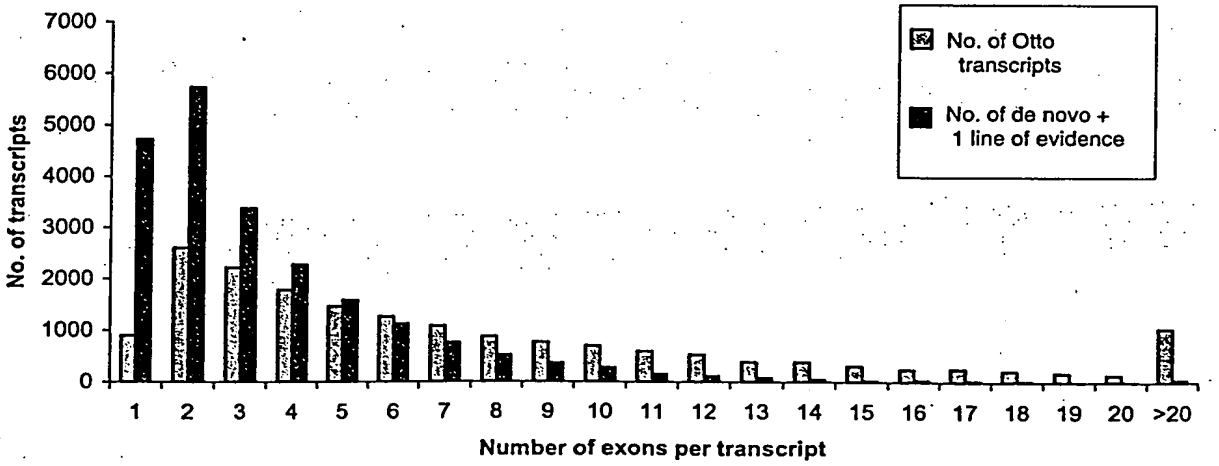


Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

examined. Unfortunately, too few meiotic crossovers have occurred in Centre d'Étude du Polymorphism Humain (CEPH) and other reference families to provide a resolution any finer than about 3 Mbp. The next challenge will be to determine a sequence basis of recombination at the chromosomal level. An accurate predictor for the rate for variation in recombination rates between any pair of markers would be extremely useful in designing markers to narrow a region of linkage, such as in positional cloning projects.

4.3 Correlation between CpG islands and genes

CpG islands are stretches of unmethylated DNA with a higher frequency of CpG dinucleotides when compared with the entire genome (*74*). CpG islands are believed to preferentially occur at the transcriptional start of genes, and it has been observed that most housekeeping genes have CpG islands at the 5′ end of the transcript (*75*, *76*). In addition, experimental evidence indicates that CpG island methylation is correlated with gene inactivation (*77*) and has been shown to be important during gene imprinting (*78*) and tissue-specific gene expression (*79*)

Experimental methods have been used that resulted in an estimate of 30,000 to 45,000 CpG islands in the human genome (*74*, *80*) and an estimate of 499 CpG islands on human chromosome 22 (*81*). Larsen *et al.* (*76*) and Gardiner-Garden and Frommer (*75*) used a computational method to identify CpG islands and defined them as regions of DNA of >200 bp that have a G+C content of >50% and a ratio of observed versus expected frequency of CG dinucleotide ≥0.6.

It is difficult to make a direct comparison of experimental definitions of CpG islands with computational definitions because computational methods do not consider the methylation state of cytosine and experimental methods do not directly select regions of high G+C content. However, we can determine the correlation of CpG island with gene starts, given a set of annotated genomic transcripts and the whole genome sequence. We have analyzed the publicly available annotation of chromosome 22, as well as using the entire human genome in our assembly and the computationally annotated genes. A variation of the CpG island computation was compared with Larsen *et al.* (*76*). The main differences are that we use a sliding window of 200 bp, consecutive windows are merged only if they overlap, and we recompute the CpG value upon merging, thus rejecting any potential island if it scores less than the threshold.

To compute various CpG statistics, we used two different thresholds of CG dinucleotide likelihood ratio. Besides using the original threshold of 0.6 (method 1), we used a higher threshold of CG dinucleotide likelihood ratio of 0.8 (method 2), which results in the number of CpG islands on chromosome 22 close to the number of annotated genes on this chromosome. The main results are summarized in Table 13. CpG islands computed with method 1 predicted only 2.6% of the CSA sequence as CpG, but 40% of the gene starts (start codons) are contained inside a CpG island. This is comparable to ratios reported by others (*82*). The last two rows of the table show the observed and expected average distance, respectively, of the closest CpG island from the first exon. The observed average closest CpG islands are smaller than the corresponding expected distances, confirming an association between CpG island and the first exon.

We also looked at the distribution of CpG island nucleotides among various sequence classes such as intergenic regions, introns, exons, and first exons. We computed the likelihood score for each sequence class as the ratio of the observed fraction of CpG island nucleotides in that sequence class and the expected fraction of CpG island nucleotides in that sequence class. The result of applying method 1 on CSA were scores of 0.89 for intergenic region, 1.2 for intron, 5.86 for exon, and 13.2 for first exon. The same trend was also found for chromosome 22 and after the application of a higher threshold (method 2) on both data sets. In sum, genome-wide analysis has extended earlier analysis and suggests a strong correlation between CpG islands and first coding exons.

4.4 Genome-wide repetitive elements

The proportion of the genome covered by various classes of repetitive DNA is presented in Table 14. We observed about 35% of the genome in these repeat classes, very similar to values reported previously (*83*). Repetitive sequence may be underrepresented in the Celera assembly as a result of incomplete repeat resolution, as discussed above. About 8% of the scaffold length is in gaps, and we expect that much of this is repetitive sequence. Chromosome 19 has the highest repeat density (57%), as well as the highest gene density (Table 10). Of interest, among the different classes of repeat elements, we observe a clear association of Alu elements and gene density, which was not observed between LINEs and gene density.

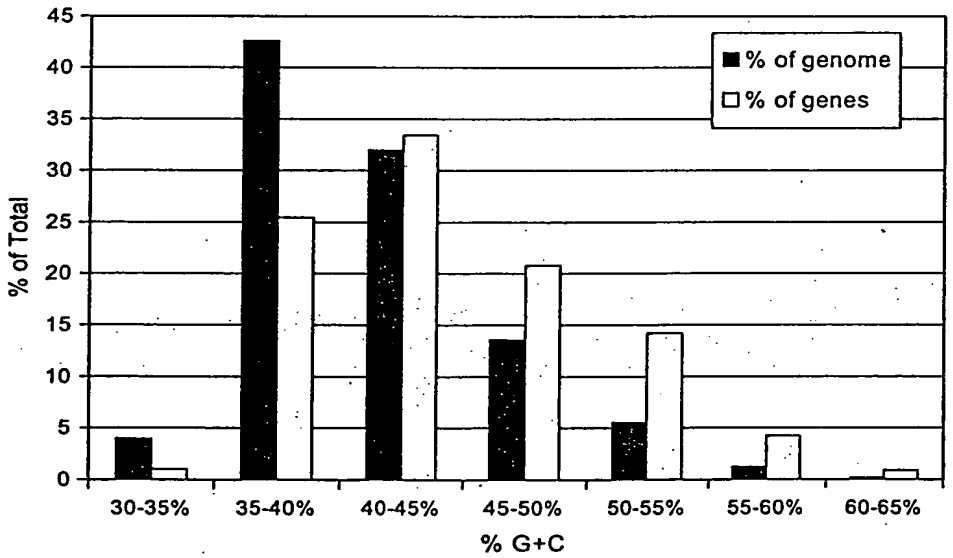


Fig. 10. Relation between G+C content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated G+C content. The percent of the total number of genes associated with each G+C bin is represented by the yellow bars. The graph shows that about 5% of the genome has a G+C content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

5 Genome Evolution

Summary. The dynamic nature of genome evolution can be captured at several levels. These include gene duplications mediated by RNA intermediates (retrotransposition) and segmental genomic duplications. In this section, we document the genome-wide occurrence of retrotransposition events generating functional (intronless paralogs) or inactive genes (pseudogenes). Genes involved in translational processes and nuclear regulation account for nearly 50% of all intronless paralogs and processed pseudogenes detected in our survey. We have also cataloged the extent of segmental genomic duplication and provide evidence for 1077 duplicated blocks covering 3522 distinct genes.

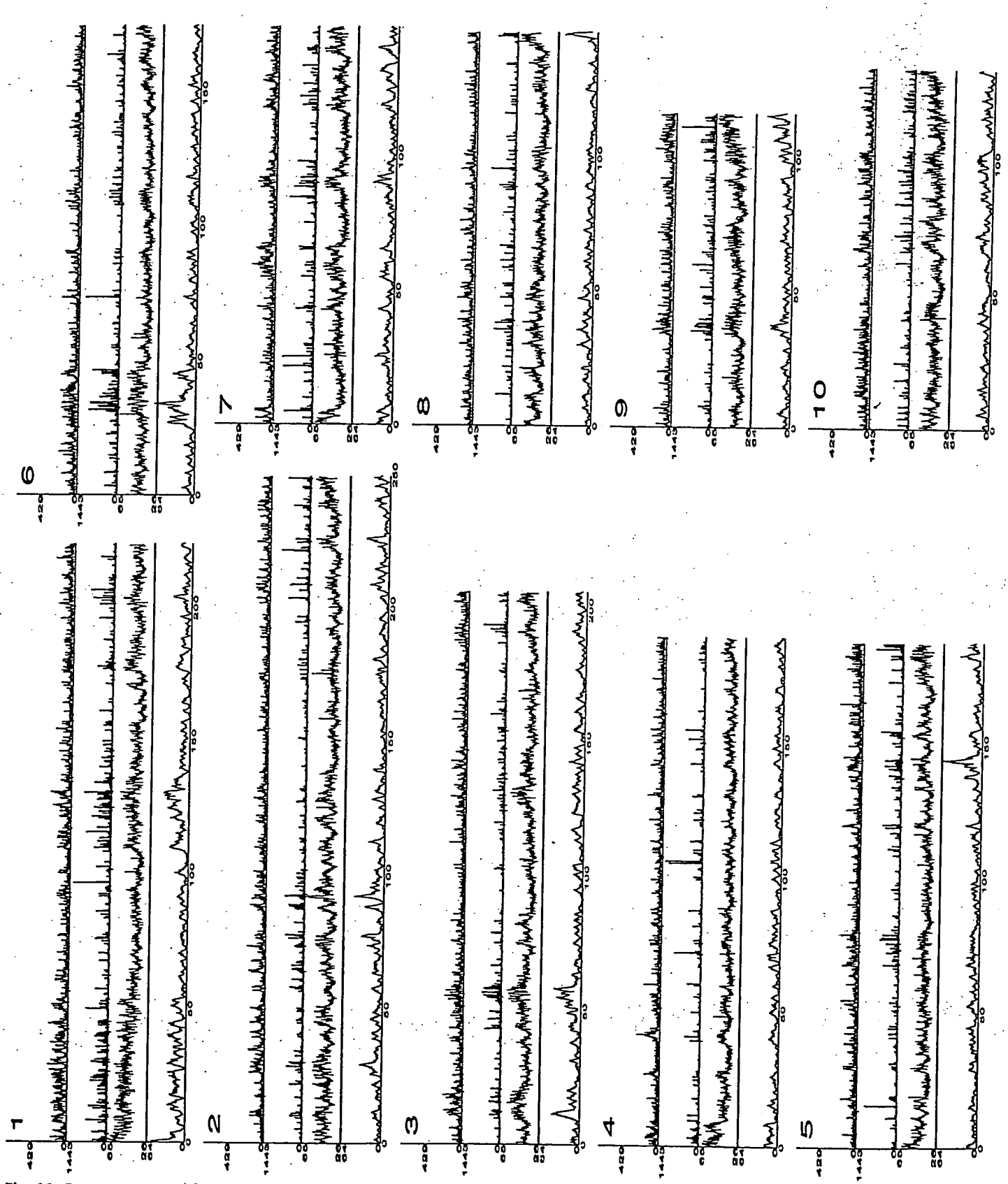


Fig. 11. Genome structural features.

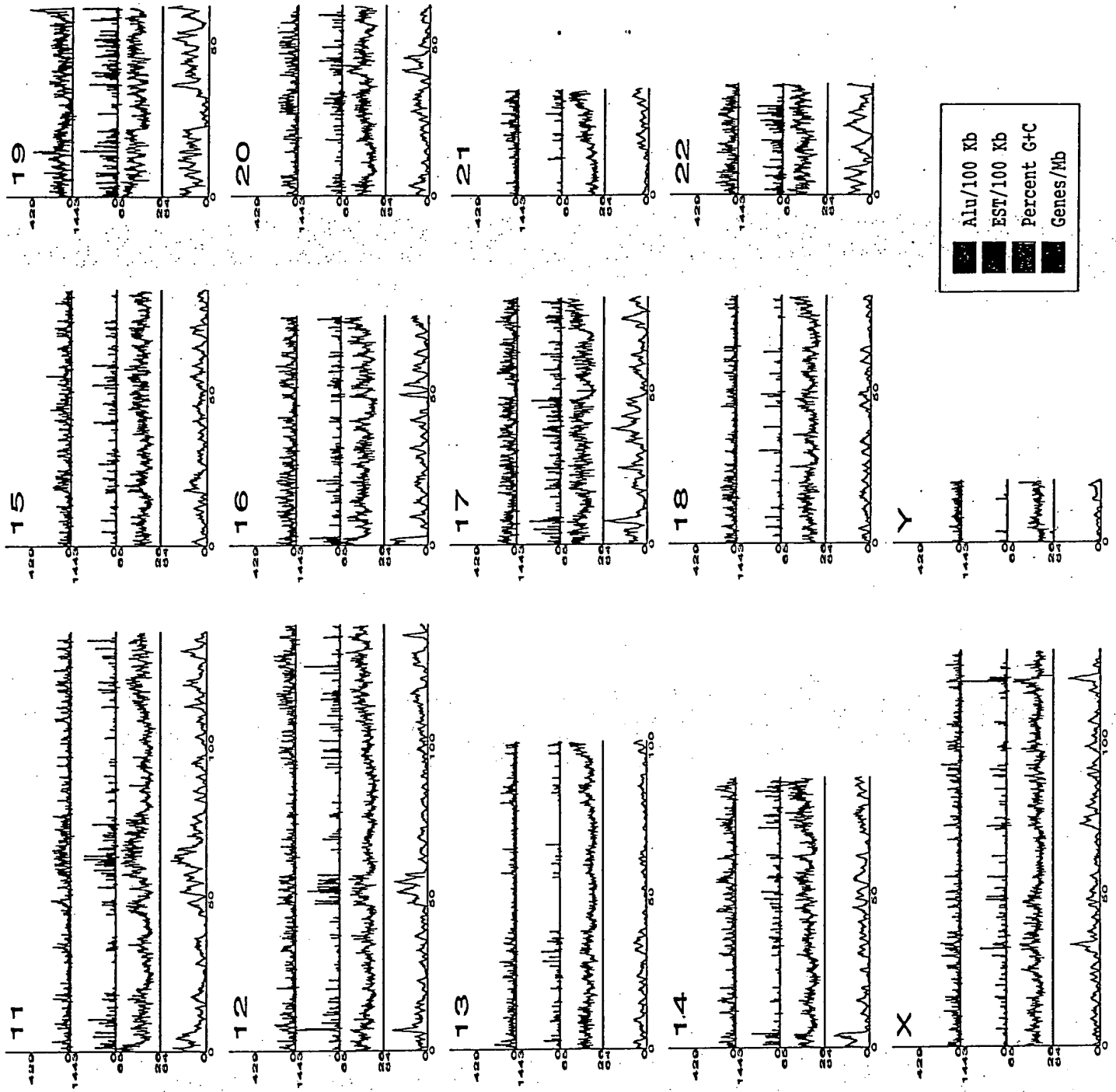


Fig. 11 (continued). Relation among gene density (orange), G+C content (green), EST density (blue), and Alu density (pink) along the lengths of each of the chromosomes. Gene density was calculated in 1-Mbp windows. The percent of G+C nucleotides was calculated in 100-kbp windows. The number of ESTs and Alu elements is shown per 100-kbp window.

5.1 Retrotransposition in the human genome

Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of a duplication event. The existence of both intron-containing and intronless forms of genes encoding functionally similar or identical proteins has been previously described (*84*, *85*). Cataloging these evolutionary events on the genomic landscape is of value in understanding the functional consequences of such gene-duplication events in cellular biology. Identification of conserved intronless paralogs in the mouse or other mammalian genomes should provide the basis for capturing the evolutionary chronology of these transposition events and provide insights into gene loss and accretion in the mammalian radiation.

A set of proteins corresponding to all 901

Table 10. Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2x refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaffolds	Largest scaffold (Mbp)	No. of scaffolds >500 kbp	Sequence covered by scaffolds >500 kbp	% of total sequence in scaffolds >500 kbp	% repeat	% GC	No of CpG Islands	Otto	De novo/ any	De novo/ 2X	Total (Otto + de novo/ any)	Total (Otto + de novo/ any)	Sequence in deserts >500/ kbp	Sequence in deserts >1 Mbp	Otto	De novo/ any	De novo/ 2X	Otto + de novo/ any	Otto + de novo/ 2X
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

*Chromosomal assignment unknown.

Otto-predicted, single-exon genes were subjected to BLAST analysis against the proteins encoded by the remaining multiexon predicted transcripts. Using homology criteria of 70% sequence identity over 90% of the length, we identified 298 instances of single- to multi-exon correspondence. Of these 298 sequences, 97 were represented in the GenBank data set of experimentally validated full-length genes at the stringency specified and were verified by manual inspection.

We believe that these 97 cases may represent intronless paralogs (see Web table 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) of known genes. Most of these are flanked by direct repeat sequences, although the precise nature of these repeats remains to be determined. All of the cases for which we have high confidence contain polyadenylated [poly(A)] tails characteristic of retrotransposition.

Recent publications describing the phenomenon of functional intronless paralogs speculate that retrotransposition may serve as a mechanism used to escape X-chromosomal inactivation (*84*, *86*). We do not find a bias toward X chromosome origination of these retrotransposed genes; rather, the results show a random chromosome distribution of both the intron-containing and corresponding intronless paralogs. We also have found several cases of retrotransposition from a single source chromosome to multiple target chromosomes. Interesting examples include the retrotransposition of a five exon–containing ribosomal protein L21 gene on chromosome 13 onto chromosomes 1, 3, 4, 7, 10, and 14, respectively. The size of the source genes can also show variability. The largest example is the 31-exon diacylglycerol kinase zeta gene on chromosome 11 that has an intronless paralog on chromosome 13. Regardless of route, retrotransposition with subsequent gene changes in coding or noncoding regions that lead to different functions or expression patterns, represents a key route to providing an enhanced functional repertoire in mammals (*87*).

Our preliminary set of retrotransposed intronless paralogs contains a clear overrepresentation of genes involved in translational processes (40% ribosomal proteins and 10% translation elongation factors) and nuclear regulation (HMG nonhistone proteins, 4%), as well as metabolic and regulatory enzymes. EST matches specific to a subset of intronless paralogs suggest expression of these intronless paralogs. Differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated will require further elucidation and experimental validation.

5.2 Pseudogenes

A pseudogene is a nonfunctional copy that is very similar to a normal gene but that has been altered slightly so that it is not expressed. We developed a method for the preliminary analysis of processed pseudogenes in the human genome as a starting point in elucidating the ongoing evolutionary forces

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

*In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Table 12. Rate of recombination per physical distance (cM/Mb) across the genome. Genethon markers were placed on CSA-mapped assemblies, and then relative physical distances and rates were calculated in 3-Mb windows for each chromosome. NA, not applicable.

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

that account for gene inactivation. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly(A) tract at the 3′ end, and direct repeats flanking the pseudogene sequence. Processed pseudogenes occur as a result of retrotransposition, whereas unprocessed pseudogenes arise from segmental genome duplication.

We searched the complete set of Otto-predicted transcripts against the genomic sequence by means of BLAST. Genomic regions corresponding to all Otto-predicted transcripts were excluded from this analysis. We identified 2909 regions matching with greater than 70% identity over at least 70% of the length of the transcripts that likely represent processed pseudogenes. This number is probably an underestimate because specific methods to search for pseudogenes were not used.

We looked for correlations between structural elements and the propensity for retrotransposition in the human genome. GC content and transcript length were compared between the genes with processed pseudogenes (1177 source genes) versus the remainder of the predicted gene set. Transcripts that give rise to processed pseudogenes have shorter average transcript length (1027 bp versus 1594 bp for the Otto set) as compared with genes for which no pseudogene was detected. The overall GC content did not show any significant difference, contrary to a recent report (*88*). There is a clear trend in gene families that are present as processed pseudogenes. These include ribosomal proteins (67%), lamin receptors (10%), translation elongation factor alpha (5%), and HMG–non-histone proteins (2%). The increased occurrence of retrotransposition (both intronless paralogs and processed pseudogenes) among genes involved in translation and nuclear regulation may reflect an increased transcriptional activity of these genes.

5.3 Gene duplication in the human genome

Building on a previously published procedure (*27*), we developed a graph-theoretic algorithm, called Lek, for grouping the predicted human protein set into protein families (*89*). The complete clusters that result from the Lek clustering provide one basis for comparing the role of whole-genome or chromosomal duplication in protein family expansion as opposed to other means, such as tandem duplication. Because each complete cluster represents a closed and certain island of homology, and because Lek is capable of simultaneously clustering protein complements of several organisms, the number of proteins contributed by each organism to a complete cluster can be predicted with confidence depending on the quality of the annotation of each genome. The variance of each organism's contribution to each cluster can then be calculated, allowing an assessment of the relative importance of large-scale duplication versus smaller-scale, organism-specific expansion and contraction of protein families, presumably as a result of natural selection operating on individual protein families within an organism. As can be seen in Fig. 12, the large variance in the relative numbers of human as compared with *D. melanogaster* and *Caenorhabditis elegans* proteins in complete clusters may be explained by multiple events of relative expansions in gene families in each of the three animal genomes. Such expansions would give rise to the distribution that shows a peak at 1:1 in the ratio for human-worm or human-fly clusters with the slope spread covering both human and fly/worm predominance, as we observed (Fig. 12). Furthermore, there are nearly as many clusters where worm and fly proteins predominate despite the larger numbers of proteins in the human. At face value, this analysis suggests that natural selection acting on individual protein families has been a major force driving the expansion of at least some elements of the human protein set. However, in our analysis, the difference between an ancient whole-genome duplication followed by loss, versus piecemeal duplication, cannot be easily distinguished. In order to differentiate these scenarios, more extended analyses were performed.

Table 13. Characteristics of CpG islands identified in chromosome 22 (34-Mbp sequence length) and the whole genome (2.9-Gbp sequence length) by means of two different methods. Method 1 uses a CG likelihood ratio of ≥0.6. Method 2 uses a CG likelihood ratio of ≥0.8.

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

Table 14. Distribution of repetitive DNA in the compartmentalized shotgun assembly sequence.

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (*83*)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

5.4 Large-scale duplications

Using two independent methods, we searched for large-scale duplications in the human genome. First, we describe a protein family–based method that identified highly conserved blocks of duplication. We then describe our comprehensive method for identifying all interchromosomal block duplications. The latter method identified a large number of duplicated chromosomal segments covering parts of all 24 chromosomes.

The first of the methods is based on the idea of searching for blocks of highly conserved homologous proteins that occur in more than one location on the genome. For this comparison, two genes were considered equivalent if their protein products were de-

termined to be in the same family and the same complete Lek cluster (essentially paralogous genes) (*89*). Initially, each chromosome was represented as a string of genes ordered by the start codons for predicted genes along the chromosome. We considered the two strands as a single string, because local inversions are relatively common events relative to large-scale duplications. Each gene was indexed according to the protein family and Lek complete cluster (*89*). All pairs of indexed gene strings were then aligned in both the forward and reverse directions with the Smith-Waterman algorithm (*90*). A match between two proteins of the same Lek complete cluster was given a score of 10 and a mismatch −10, with gap open and extend penalties of −4 and −1. With these parameters, 19 conserved interchromosomal blocks of duplication were observed, all of which were also detected and expanded by the comprehensive method described below. The detection of only a relatively small number of block duplications was a consequence of using an intrinsically conservative method grounded in the conservative constraints of the complete Lek clusters.

In the second, more comprehensive approach, we aligned all chromosomes directly with one another using an algorithm based on the MUMmer system (*91*). This alignment method uses a suffix tree data structure and a linear-time algorithm to align long sequences very rapidly; for example, two chromosomes of 100 Mbp can be aligned in less than 20 min (on a Compaq Alpha computer) with 4 gigabytes of memory. This procedure was used recently to identify numerous large-scale segmental duplications among the five chromosomes of *A. thaliana* (*92*); in that organism, the method revealed that 60% of the genome (66 Mbp) is covered by 24 very large duplicated segments. For *Arabidopsis*, a DNA-based alignment was sufficient to reveal the segmental duplications between chromosomes; in the human genome, DNA alignments at the whole-chromosome level are insufficiently sensitive. Therefore, a modified procedure was developed and applied, as follows. First, all 26,588 proteins (9,675,713 million amino acids) were concatenated end-to-end in order as they occur along each of the 24 chromosomes, irrespective of strand location. The concatenated protein set was then aligned against each chromosome by the MUMmer algorithm. The resulting matches were clustered to extract all sets of three or more protein matches that occur in close proximity on two different chromosomes (*93*); these represent the candidate segmental duplications. A series of filters were developed and applied to remove likely false-positives from this set; for example, small blocks that were spread across many proteins were removed. To refine the filtering methods, a shuffled protein set was first created by taking the 26,588 proteins, randomizing their order, and then partitioning them into 24 shuffled chromosomes, each containing the same number of proteins as the true genome. This shuffled protein set has the identical composition to the real genome; in particular, every protein and every domain appears the same number of times. The complete algorithm was then applied to both the real and the shuffled data, with the results on the shuffled data being used to estimate the false-positive rate. The algorithm after filtering yielded 10,310 gene pairs in 1077 duplicated blocks containing 3522 distinct genes; tandemly duplicated expansions in many of the blocks explain the excess of gene pairs to distinct genes. In the shuffled data, by contrast, only 370 gene pairs were found, giving a false-positive estimate of 3.6%. The most likely explanation for the 1077 block duplications is ancient segmental duplications. In many cases, the order of the proteins has been shuffled, although proximity is preserved. Out of the 1077 blocks, 159 contain only three genes, 137 contain four genes, and 781 contain five or more genes.

To illustrate the extent of the detected duplications, Fig. 13 shows all 1077 block duplications indexed to each chromosome in 24 panels in which only duplications mapped to the indexed chromosome are displayed. The figure makes it clear that the duplications are ubiquitous in the genome. One feature that it displays is many relatively small chromosomal stretches, with one-to-many duplication relationships that are graphically striking. One such example captured by the analysis is the well-documented olfactory receptor (OR) family, which is scattered in blocks throughout the genome and which has been analyzed for genome-deployment reconstructions at several evolutionary stages (*94*). The figure also illustrates that some chromosomes, such as chromosome 2, contain many more detected large-scale duplications than others. Indeed, one of the largest duplicated segments is a large block of 33 proteins on chromosome 2, spread among eight smaller blocks in 2p, that aligns to a paralogous set on chromosome 14, with one rearrangement (see chromosomes 2 and 14 panels in Fig. 13). The proteins are not contiguous but span a region containing 97 proteins on chromosome 2 and 332 proteins on chromosome 14. The likelihood of observing this many duplicated proteins by chance, even over a span of this length, is 2.3×10^{-68} (*93*). This duplicated set spans 20 Mbp on chromosome 2 and 63 Mbp on chromosome 14, over 70% of the latter chromosome. Chromosome 2 also contains a block duplication that is nearly as large, which is shared by chromosome arm 2q and chromosome 12. This duplication incorporates two of the four known Hox gene clusters, but considerably expands the extent of the duplications proximally and distally on the pair of chromosome arms. This breadth of duplication is also seen on the two chromosomes carrying the other two Hox clusters.

An additional large duplication, between chromosomes 18 and 20, serves as a good example to illustrate some of the features common to many of the other observed large duplications (Fig. 13, inset). This duplication contains 64 detected ordered intrachromosomal pairs of homologous genes. After discounting a 40-Mb stretch of chromosome 18 free of matches to chromosome 20, which is likely to represent a large insert (between the gene assignments "Krup rel" and "collagen rel" on chromosome 18 in Fig. 13), the full duplication segment covers 36 Mb on chromosome 18 and 28 Mb on chromosome 20.

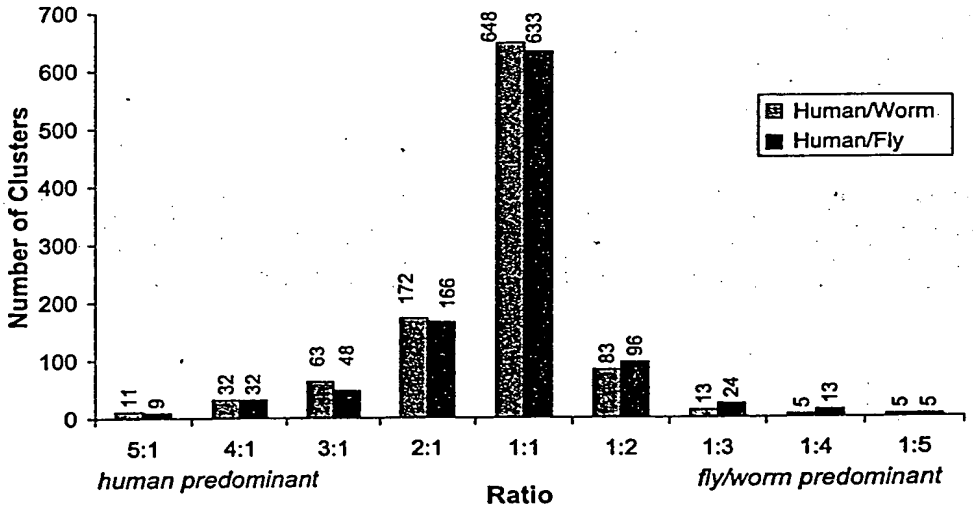


Fig. 12. Gene duplication in complete protein clusters. The predicted protein sets of human, worm, and fly were subjected to Lek clustering (*27*). The numbers of clusters with varying ratios (whole number) of human versus worm and human versus fly proteins per cluster were plotted.

By this measure, the duplication segment spans nearly half of each chromosome's net length. The most likely scenario is that the whole span of this region was duplicated as a single very large block, followed by shuffling owing to smaller scale rearrangements. As such, at least four subsequent rearrangements would need to be invoked to explain the relative insertions and inversions seen in the duplicated segment interval. The 64 protein pairs in this alignment occur among 217 protein assignments on chromosome 18, and among 322 protein assignments on chromosome 20, for a density of involved proteins of 20 to 30%. This is consistent with an ancient large-scale duplication followed by subsequent gene loss on one or both chromosomes. Loss of just one member of a gene pair subsequent to the duplication would result in a failure to score a gene pair in the block; less than 50% gene loss on the chromosomes would lead to the duplication density observed here. As an independent verification of the significance of the alignments detected, it can be seen that a substantial number of the pairs of aligning proteins in this duplication, including some of those annotated (Fig. 13), are those populating small Lek complete clusters (see above). This indicates that they are members of very small families of paralogs; their relative scarcity within the genome validates the uniqueness and robust nature of their alignments.

Two additional qualitative features were observed among many of the large-scale duplications. First, several proteins with disease associations, with OMIM (Online Mendelian Inheritance in Man) assignments, are members of duplicated segments (see web table 2 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1). We have also observed a few instances where paralogs on both duplicated segments are associated with similar disease conditions. Notable among these genes are proteins involved in hemostasis (coagulation factors) that are associated with bleeding disorders, transcriptional regulators like the homeobox proteins associated with developmental disorders, and potassium channels associated with cardiovascular conduction abnormalities. For each of these disease genes, closer study of the paralogous genes in the duplicated segment may reveal new insights into disease causation, with further investigation needed to determine whether they might be involved in the same or similar genetic diseases. Second, although there is a conserved number of proteins and coding exons predicted for specific large duplicated spans within the chromosome 18 to 20 alignment, the genomic DNA of chromosome 18 in these specific spans is in some cases more than 10-fold longer than the corresponding chromosome 20 DNA. This selective accretion of noncoding DNA (or conversely, loss of noncoding DNA) on one of a pair of duplicated chromosome regions was observed in many compared regions. Hypotheses to explain which mechanisms foster these processes must be tested.

Evaluation of the alignment results gives some perspective on dating of the duplications. As noted above, large-scale ancient segmental duplication in fact best explains many of the blocks detected by this genome-wide analysis. The regions of human chromosomes involved in the large-scale duplications expanded upon above (chromosomes 2 to 14, 2 to 12, and 18 to 20) are each syntenic to a distinct mouse chromosomal region. The corresponding mouse chromosomal regions are much more similar in sequence conservation, and even in order, to their human synteny partners than the human duplication regions are to each other. Further, the corresponding mouse chromosomal regions each bear a significant proportion of genes orthologous to the human genes on which the human duplication assignments were made. On the basis of these factors, the corresponding mouse chromosomal spans, at coarse resolution, appear to be products of the same large-scale duplications observed in humans. Although further detailed analysis must be carried out once a more complete genome is assembled for mouse, the underlying large duplications appear to predate the two species' divergence. This dates the duplications, at the latest, before divergence of the primate and rodent lineages. This date can be further refined upon examination of the synteny between human chromosomes and those of chicken, pufferfish (*Fugu rubripes*), or zebrafish (*95*). The only substantial syntenic stretches mapped in these species corresponding to both pairs of human duplications are restricted to the Hox cluster regions. When the synteny of these regions (or others) to human chromosomes is extended with further mapping, the ages of the nearly chromosome-length duplications seen in humans are likely to be dated to the root of vertebrate divergence.

The MUMmer-based results demonstrate large block duplications that range in size from a few genes to segments covering most of a chromosome. The extent of segmental duplications raises the question of whether an ancient whole-genome duplication event is the underlying explanation for the numerous duplicated regions (*96*). The duplications have undergone many deletions and subsequent rearrangements; these events make it difficult to distinguish between a whole-genome duplication and multiple smaller events. Further analysis, focused especially on comparing the estimated ages of all the block duplications, derived partially from interspecies genome comparisons, will be necessary to determine which of these two hypotheses is more likely. Comparisons of genomes of different vertebrates, and even cross-phyla genome comparisons, will allow for the deconvolution of duplications to eventually reveal the stagewise history of our genome, and with it a history of the emergence of many of the key functions that distinguish us from other living things.

6 A Genome-Wide Examination of Sequence Variations

Summary. Computational methods were used to identify single-nucleotide polymorphisms (SNPs) by comparison of the Celera sequence to other SNP resources. The SNP rate between two chromosomes was ~1 per 1200 to 1500 bp. SNPs are distributed nonrandomly throughout the genome. Only a very small proportion of all SNPs (<1%) potentially impact protein function based on the functional analysis of SNPs that affect the predicted coding regions. This results in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins.

Having a complete genome sequence enables researchers to achieve a dramatic acceleration in the rate of gene discovery, but only through analysis of sequence variation in DNA can we discover the genetic basis for variation in health among human beings. Whole-genome shotgun sequencing is a particularly effective method for detecting sequence variation in tandem with whole-genome assembly. In addition, we compared the distribution and attributes of SNPs ascertained by three other methods: (i) alignment of the Celera consensus sequence to the PFP assembly, (ii) overlap of high-quality reads of genomic sequence (referred to as "Kwok"; 1,120,195 SNPs) (*97*), and (iii) reduced representation shotgun sequencing (referred to as "TSC"; 632,640 SNPs) (*98*). These data were consistent in showing an overall nucleotide diversity of $\sim 8 \times 10^{-4}$, marked heterogeneity across the genome in SNP density, and an overwhelming preponderance of noncoding variation that produces no change in expressed proteins.

6.1 SNPs found by aligning the Celera consensus to the PFP assembly

Ideally, methods of SNP discovery make full use of sequence depth and quality at every site, and quantitatively control the rate of false-positive and false-negative calls with an explicit sampling model (*99*). Comparison of consensus sequences in the absence of these details necessitated a more ad hoc approach (quality scores could not readily be obtained for the PFP assembly). First, all sequence differences between the two consensus sequences were identified; these were then filtered to reduce the contribution of sequencing errors and misassembly. As a measure of the effectiveness of the filtering step, we monitored the ratio of transition and transversion substitutions, because a 2:1 ratio has been well documented as typical in mammalian evolution (*100*) and in human SNPs

(*101*, *102*). The filtering steps consisted of removing variants where the quality score in the Celera consensus was less than 30 and where the density of variants was greater than 5 in 400 bp. These filters resulted in shifting the transition-to-transversion ratio from 1.57:1 to 1.89:1. When applied to 2.3 Gbp of alignments between the Celera and PFP consensus sequences, these filters resulted in identification of 2,104,820 putative SNPs from a total of 2,778,474 substitution differences. Overlaps between this set of SNPs and those found by other methods are described below.

6.2 Comparisons to public SNP databases

Additional SNPs, including 2,536,021 from dbSNP (www.ncbi.nlm.nih.gov/SNP) and 13,150 from HGMD (Human Gene Mutation Database, from the University of Wales, UK), were mapped on the Celera consensus sequence by a sequence similarity search with the program PowerBlast (*103*). The two largest data sets in dbSNP are the Kwok and TSC sets, with 47% and 25% of the dbSNP records. Low-quality alignments with partial coverage of the dbSNP sequence and alignments that had less than 98% sequence identity between the Celera sequence and the dbSNP flanking sequence were eliminated. dbSNP sequences mapping to multiple locations on the Celera genome were discarded. A total of 2,336,935 dbSNP variants were mapped to 1,223,038 unique locations on the Celera sequence, implying considerable redundancy in dbSNP. SNPs in the TSC set mapped to 585,811 unique genomic locations, and SNPs in the Kwok set mapped to 438,032 unique locations. The combined unique SNPs counts used in this analysis, including Celera-PFP, TSC, and Kwok, is 2,737,668. Table 15 shows that a substantial fraction of SNPs identified by one of these methods was also found by another method. The very high overlap (36.2%) between the Kwok and Celera-PFP SNPs may be due in part to the use by Kwok of sequences that went into the PFP assembly. The unusually low overlap (16.4%) between the Kwok and TSC sets is due to their being the smallest two sets. In addition, 24.5% of the Celera-PFP SNPs overlap with SNPs derived from the Celera genome sequences (*46*). SNP validation in population samples is an expensive and laborious process, so confirmation on multiple data sets may provide an efficient initial validation "in silico" (by computational analysis).

One means of assessing whether the three sets of SNPs provide the same picture of human variation is to tally the frequencies of the six possible base changes in each set of SNPs (Table 16). Previous measures of nucleotide diversity were mostly derived from small-scale analysis on candidate genes (*101*), and our analysis with all three data sets validates the previous observations at the whole-genome scale. There is remarkable homogeneity between the SNPs found in the Kwok set, the TSC set, and in our whole-genome shotgun (*46*) in this substitution pattern. Compared with the rest of the data sets, Celera-PFP deviates slightly from the 2:1 transition-to-transversion ratio observed in the other SNP sets. This result is not unexpected, because some fraction of the computationally identified SNPs in the Celera-PFP comparison may in fact be sequence errors. A 2:1 transition:transversion ratio for the bona fide SNPs would be obtained if one assumed that 15% of the sequence differences in the Celera-PFP set were a result of (presumably random) sequence errors.

6.3 Estimation of nucleotide diversity from ascertained SNPs

The number of SNPs identified varied widely across chromosomes. In order to normalize these values to the chromosome size and sequence coverage, we used π, the standard statistic for nucleotide diversity (*104*). Nucleotide diversity is a measure of per-site heterozygosity, quantifying the probability that a pair of chromosomes drawn from the population will differ at a nucleotide site. In order to calculate nucleotide diversity for each chromosome, we need to know the number of nucleotide sites that were surveyed for variation, and in methods like reduced respresentation sequencing, we need to know the sequence quality and the depth of coverage at each site. These data are not readily available, so we could not estimate nucleotide diversity from the TSC effort. Estimation of nucleotide diversity from high-quality sequence overlaps should be possible, but again, more information is needed on the details of all the alignments.

Estimation of nucleotide diversity from a shotgun assembly entails calculating for each column of the multialignment, the probability that two or more distinct alleles are present, and the probability of detecting a SNP if in fact the alleles have different sequence (i.e., the probability of correct sequence calls). The greater the depth of coverage and the higher the sequence quality, the higher is the chance of successfully detecting a SNP (*105*). Even after correcting for variation in coverage, the nucleotide diversity appeared to vary across autosomes. The significance of this heterogeneity was tested by analysis of variance, with estimates of π for 100-kbp windows to estimate variability within chromosomes (for the Celera-PFP comparison, $F = 29.73$, $P < 0.0001$).

Average diversity for the autosomes estimated from the Celera-PFP comparison was 8.94×10^{-4}. Nucleotide diversity on the X chromosome was 6.54×10^{-4}. The X is expected to be less variable than autosomes, because for every four copies of autosomes in the population, there are only three X chromosomes, and this smaller effective population size means that random drift will more rapidly remove variation from the X (*106*).

Having ascertained nucleotide variation genome-wide, it appears that previous estimates of nucleotide diversity in humans based on samples of genes were reasonably accurate (*101*, *102*, *106*, *107*). Genome-wide, our estimate of nucleotide diversity was 8.98×10^{-4} for the Celera-PFP alignment, and a published estimate averaged over 10 densely resequenced human genes was 8.00×10^{-4} (*108*).

6.4 Variation in nucleotide diversity across the human genome

Such an apparently high degree of variability among chromosomes in SNP density raises the question of whether there is heterogeneity at a finer scale within chromo-

Table 15. Overlap of SNPs from genome-wide SNP databases. Table entries are SNP counts for each pair of data sets. Numbers in parentheses are the fraction of overlap, calculated as the count of overlapping SNPs divided by the number of SNPs in the smaller of the two databases compared. Total SNP counts for the databases are: Celera-PFP, 2,104,820; TSC, 585,811; and Kwok 438,032. Only unique SNPs in the TSC and Kwok data sets were included.

	TSC	Kwok
Celera-PFP	188,694 (0.322)	158,532 (0.362)
TSC		72,024 (0.164)

Table 16. Summary of nucleotide changes in different SNP data sets.

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

*November 2000 release of the NCBI database dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pui-Yan Kwok from Washington University. †November 2000 release of NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

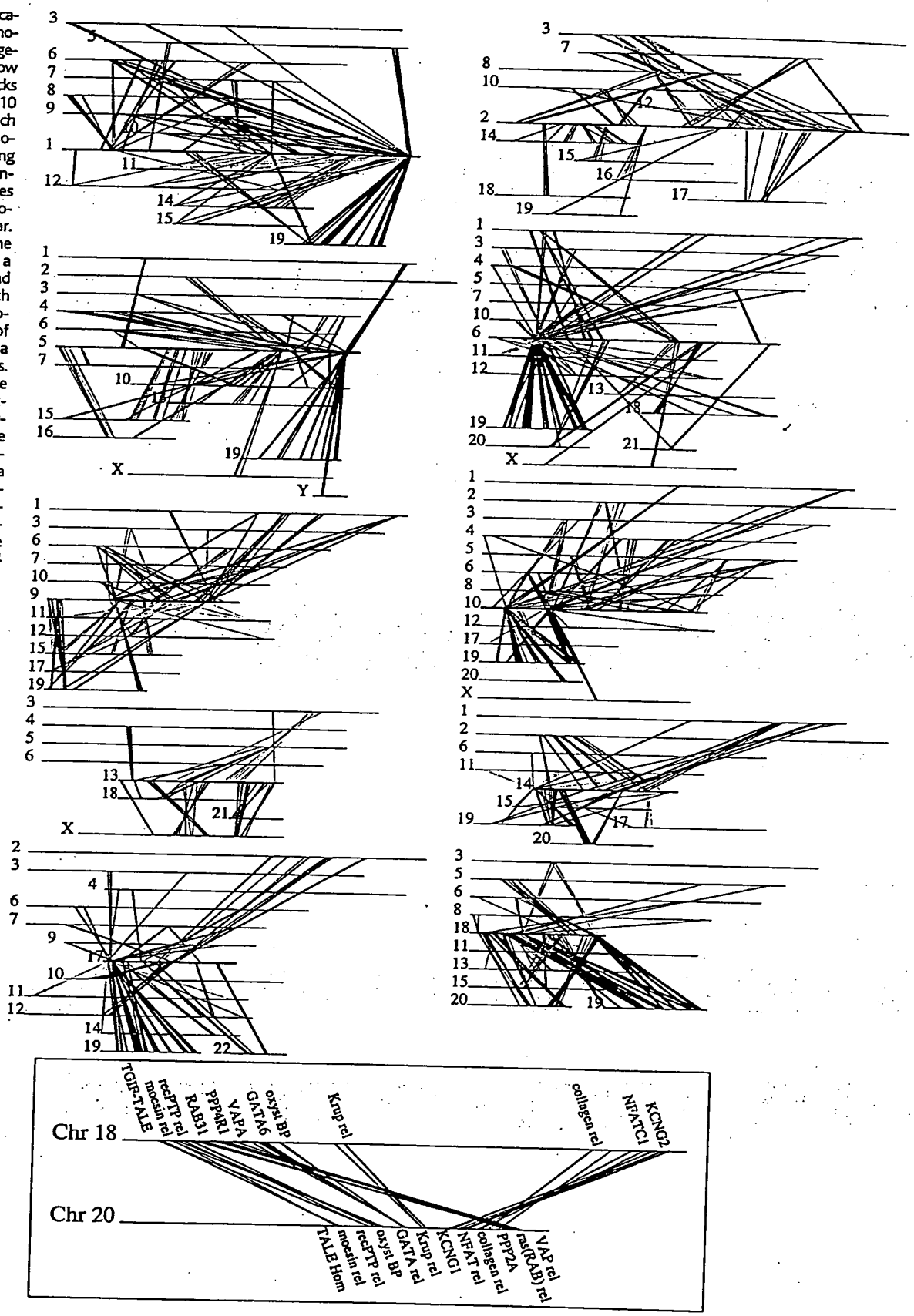
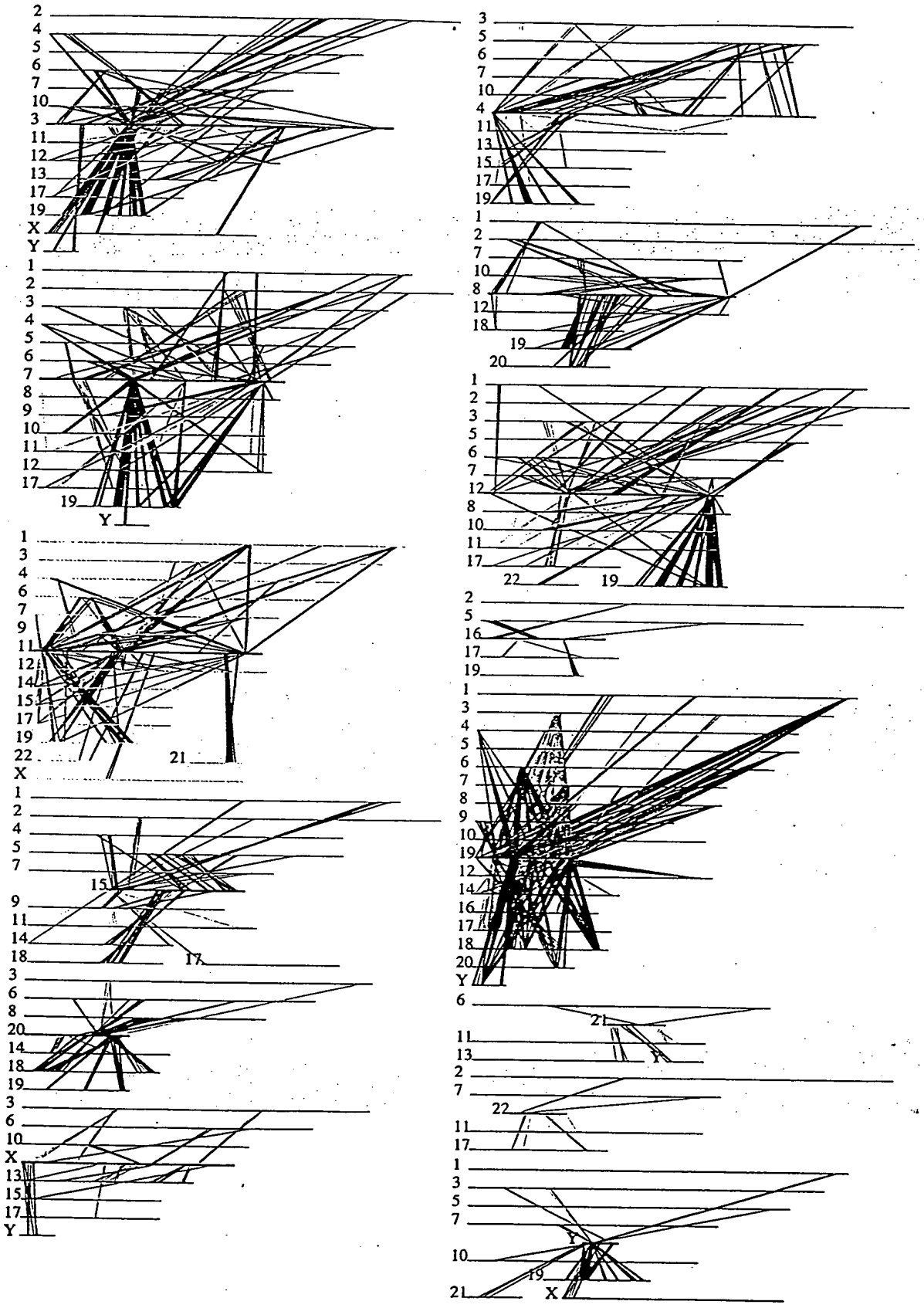


Fig. 13. Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.



somes, and whether this heterogeneity is greater than expected by chance. If SNPs occur by random and independent mutations, then it would seem that there ought to be a Poisson distribution of numbers of SNPs in fragments of arbitrary constant size. The observed dispersion in the distribution of SNPs in 100-kbp fragments was far greater than predicted from a Poisson distribution (Fig. 14). However, this simplistic model ignores the different recombination rates and population histories that exist in different regions of the genome. Population genetics theory holds that we can account for this variation with a mathematical formulation called the neutral coalescent (*109*). Applying well-tested algorithms for simulating the neutral coalescent with recombination (*110*), and using an effective population size of 10,000 and a per-base recombination rate equal to the mutation rate (*111*), we generated a distribution of numbers of SNPs by this model as well (*112*). The observed distribution of SNPs has a much larger variance than either the Poisson model or the coalescent model, and the difference is highly significant. This implies that there is significant variability across the genome in SNP density, an observation that begs an explanation.

Several attributes of the DNA sequence may affect the local density of SNPs, including the rate at which DNA polymerase makes errors and the efficacy of mismatch repair. One key factor that is likely to be associated with SNP density is the G+C content, in part because methylated cytosines in CpG dinucleotides tend to undergo deamination to form thymine, accounting for a nearly 10-fold increase in the mutation rate of CpGs over other dinucleotides. We tallied the GC content and nucleotide diversities in 100-kbp windows across the entire genome and found that the correlation between them was positive ($r = 0.21$) and highly significant ($P < 0.0001$), but G+C content accounted for only a small part of the variation.

6.5 SNPs by genomic class

To test homogeneity of SNP densities across functional classes, we partitioned sites into intergenic (defined as >5 kbp from any predicted transcription unit), 5′-UTR, exonic (missense and silent), intronic, and 3′-UTR for 10,239 known genes, derived from the NCBI RefSeq database and all human genes predicted from the Celera Otto annotation. In coding regions, SNPs were categorized as either silent, for those that do not change amino acid sequence, or missense, for those that change the protein product. The ratio of missense to silent coding SNPs in Celera-PFP, TSC, and Kwok sets (1.12, 0.91, and 0.78, respectively) shows a markedly reduced frequency of missense variants compared with the neutral expectation, consistent with the elimination by natural selection of a fraction of the deleterious amino acid changes (*112*). These ratios are comparable to the missense-to-silent ratios of 0.88 and 1.17 found by Cargill *et al.* (*101*) and by Halushka *et al.* (*102*). Similar results were observed in SNPs derived from Celera shotgun sequences (*46*).

It is striking how small is the fraction of SNPs that lead to potentially dysfunctional alterations in proteins. In the 10,239 RefSeq genes, missense SNPs were only about 0.12, 0.14, and 0.17% of the total SNP counts in Celera-PFP, TSC, and Kwok SNPs, respectively. Nonconservative protein changes constitute an even smaller fraction of missense SNPs (47, 41, and 40% in Celera-PFP, Kwok, and TSC). Intergenic regions have been virtually unstudied (*113*), and we note that 75% of the SNPs we identified were intergenic (Table 17). The SNP rate was highest in introns and lowest in exons. The SNP rate was lower in intergenic regions than in introns, providing one of the first discriminators between these two classes of DNA. These SNP rates were confirmed in the Celera SNPs, which also exhibited a lower rate in exons than in introns, and in extragenic regions than in introns (*46*). Many of these intergenic SNPs will provide valuable information in the form of markers for linkage and association studies, and some fraction is likely to have a regulatory function as well.

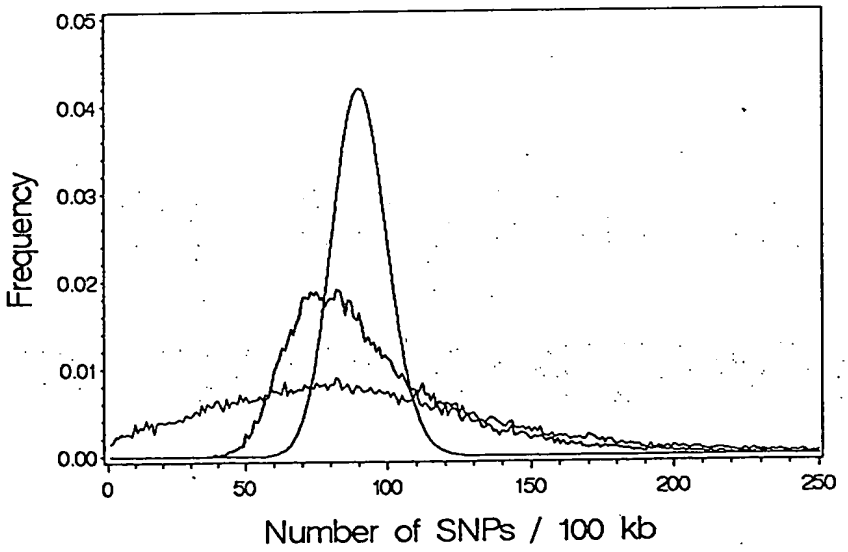


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

7 An Overview of the Predicted Protein-Coding Genes in the Human Genome

Summary. This section provides an initial computational analysis of the predicted protein set with the aim of cataloging prominent differences and similarities when the human genome is compared with other fully sequenced eukaryotic genomes. Over 40% of the predicted protein set in humans cannot be ascribed a molecular function by methods that assign proteins to known families. A protein domain–based analysis provides a detailed catalog of the prominent differences in the human genome when compared with the fly and worm genomes. Prominent among these are domain expansions in proteins involved in developmental regulation and in cellular processes such as neuronal function, hemostasis, acquired immune response, and cytoskeletal complexity. The final enumeration of protein families and details of protein structure will rely on additional experimental work and comprehensive manual curation.

A preliminary analysis of the predicted human protein-coding genes was conducted. Two methods were used to analyze and classify the molecular functions of 26,588 predicted proteins that represent 26,383 gene predictions with at least two lines of evidence as described above. The first method was based on an analysis at the level of protein families, with both the publicly available Pfam database (*114*, *115*) and Celera's Panther Classification (CPC) (Fig. 15) (*116*). The second method was based on an analysis at the level of protein domains, with both the Pfam and SMART databases (*115*, *117*).

The results presented here are preliminary and are subject to several limitations.

Both the gene predictions and functional assignments have been made by using computational tools, although the statistical models in Panther, Pfam, and SMART have been built, annotated, and reviewed by expert biologists. In the set of computationally predicted genes, we expect both false-positive predictions (some of these may in fact be inactive pseudogenes) and false-negative predictions (some human genes will not be computationally predicted). We also expect errors in delimiting the boundaries of exons and genes. Similarly, in the automatic functional assignments, we also expect both false-positive and false-negative predictions. The functional assignment protocol focuses on protein families that tend to be found across several organisms, or on families of known human genes. Therefore, we do not assign a function to many genes that are not in large families, even if the function is known. Unless otherwise specified, all enumeration of the genes in any given family or functional category was taken from the set of 26,588 predicted proteins, which were assigned functions by using statistical score cutoffs defined for models in Panther, Pfam, and SMART.

For this initial examination of the predicted human protein set, three broad questions were asked: (i) What are the likely molecular functions of the predicted gene products, and how are these proteins categorized with current classification methods? (ii) What are the core functions that appear to be common across the animals? (iii) How does the human protein complement differ from that of other sequenced eukaryotes?

7.1 Molecular functions of predicted human proteins

Figure 15 shows an overview of the putative molecular functions of the predicted 26,588 human proteins that have at least two lines of supporting evidence. About 41% (12,809) of the gene products could not be classified from this initial analysis and are termed proteins with unknown functions. Because our automatic classification methods treat only relatively large protein families, there are a number of "unclassified" sequences that do, in fact, have a known or predicted function. For the 60% of the protein set that have automatic functional predictions, the specific protein functions have been placed into broad classes. We focus here on molecular function (rather than higher order cellular processes) in order to classify as many proteins as possible. These functional predictions are based on similarity to sequences of known function.

In our analysis of the 12,731 additional low-confidence predicted genes (those with only one piece of supporting evidence), only 636 (5%) of these additional putative genes were assigned molecular functions by the automated methods. One-third of these 636 predicted genes represented endogenous retroviral proteins, further suggesting that the majority of these unknown-function genes are not real genes. Given that most of these additional 12,095 genes appear to be unique among the genomes sequenced to date, many may simply represent false-positive gene predictions.

The most common molecular functions are the transcription factors and those involved in nucleic acid metabolism (nucleic acid enzyme). Other functions that are highly represented in the human genome are the receptors, kinases, and hydrolases. Not surprisingly, most of the hydrolases are proteases. There are also many proteins that are members of proto-oncogene families, as well as families of "select regulatory molecules": (i) proteins involved in specific steps of signal transduction such as heterotrimeric GTP-binding proteins (G proteins) and cell cycle regulators, and (ii) proteins that modulate the activity of kinases, G proteins, and phosphatases.

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

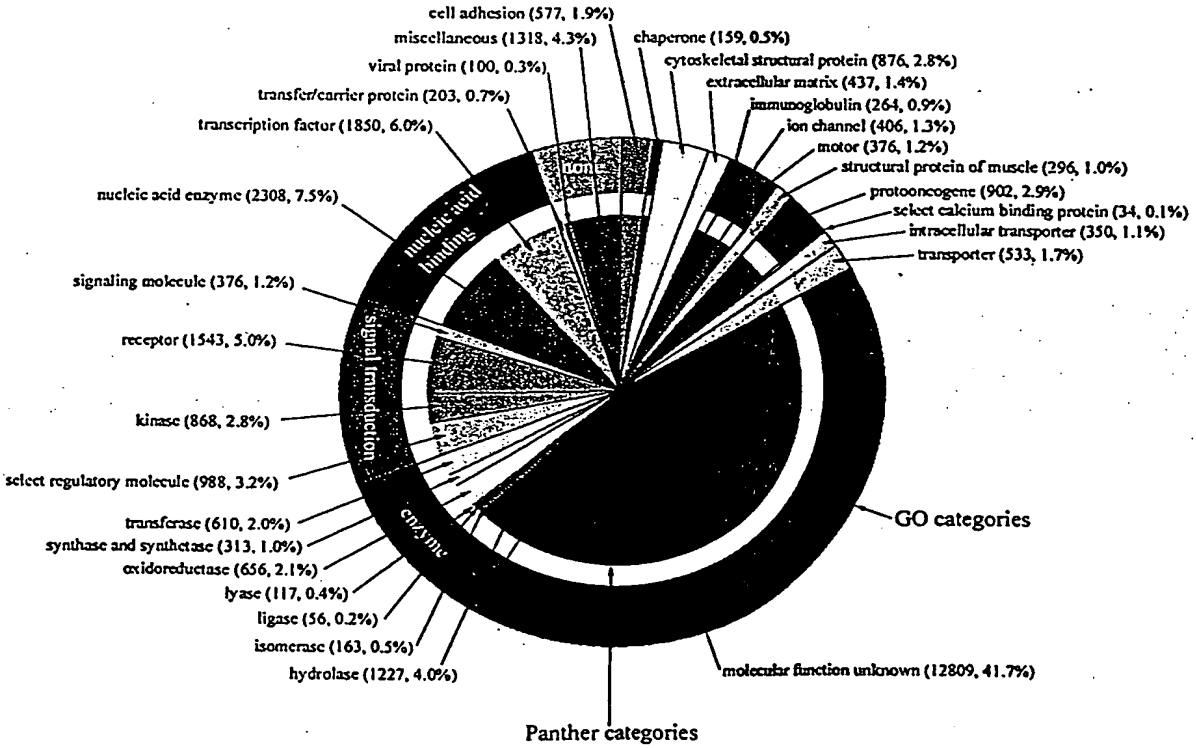


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

7.2 Evolutionary conservation of core processes

Because of the various "model organism" genome-sequencing projects that have already been completed, reasonable comparative information is available for beginning the analysis of the evolution of the human genome. The genomes of *S. cerevisiae* ("bakers' yeast") (*118*) and two diverse invertebrates, *C. elegans* (a nematode worm) (*119*) and *D. melanogaster* (fly) (*26*), as well as the first plant genome, *A. thaliana*, recently completed (*92*), provide a diverse background for genome comparisons.

We enumerated the "strict orthologs" conserved between human and fly, and between human and worm (Fig. 16) to address the question, What are the core functions that appear to be common across the animals? The concept of orthology is important because if two genes are orthologs, they can be traced by descent to the common ancestor of the two organisms (an "evolutionarily conserved protein set"), and therefore are likely to perform similar conserved functions in the different organisms. It is critical in this analysis to separate orthologs (a gene that appears in two organisms by descent from a common ancestor) from paralogs (a gene that appears in more than one copy in a given organism by a duplication event) because paralogs may subsequently diverge in function. Following the yeast-worm ortholog comparison in (*120*), we identified two different cases for each pairwise comparison (human-fly and human-worm). The first case was a pair of genes, one from each organism, for which there was no other close homolog in either organism. These are straightforwardly identified as orthologous, because there are no additional members of the families that complicate separating orthologs from paralogs. The second case is a family of genes with more than one member in either or both of the organisms being compared. Chervitz *et al.* (*120*) deal with this case by analyzing a phylogenetic tree that described the relationships between all of the sequences in both organisms, and then looked for pairs of genes that were nearest neighbors in the tree. If the nearest-neighbor pairs were from different organisms, those genes were presumed to be orthologs. We note that these nearest neighbors can often be confidently identified from pairwise sequence comparison without having to examine a phylogenetic tree (see legend to Fig. 16). If the nearest neighbors are not from different organisms, there has been a paralogous expansion in one or both organisms after the speciation event (and/or a gene loss by one organism). When this one-to-one correspondence is lost, defining an ortholog becomes ambiguous. For our initial computational overview of the predicted human protein set, we could not answer this question for every predicted protein. Therefore, we consider only "strict orthologs," i.e., the proteins with unambiguous one-to-one relationships (Fig. 16). By these criteria, there are 2758 strict human-fly orthologs, 2031 human-worm (1523 in common between these sets). We define the evolutionarily conserved set as those 1523 human proteins that have strict orthologs in both *D. melanogaster* and *C. elegans*.

The distribution of the functions of the conserved protein set is shown in Fig. 16. Comparison with Fig. 15 shows that, not surprisingly, the set of conserved proteins is not distributed among molecular functions in the same way as the whole human protein set. Compared with the whole human set (Fig. 15), there are several categories that are overrepresented in the conserved set by a factor of ~2 or more. The first category is nucleic acid enzymes, primarily the transcriptional machinery (notably DNA/RNA methyltransferases, DNA/RNA polymerases, helicases, DNA ligases, DNA- and RNA-processing factors, nucleases, and ribosomal proteins). The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals. Other enzyme types are also overrepresented (transferases, oxidoreductases, ligases, lyases, and isomerases). Many of these en-

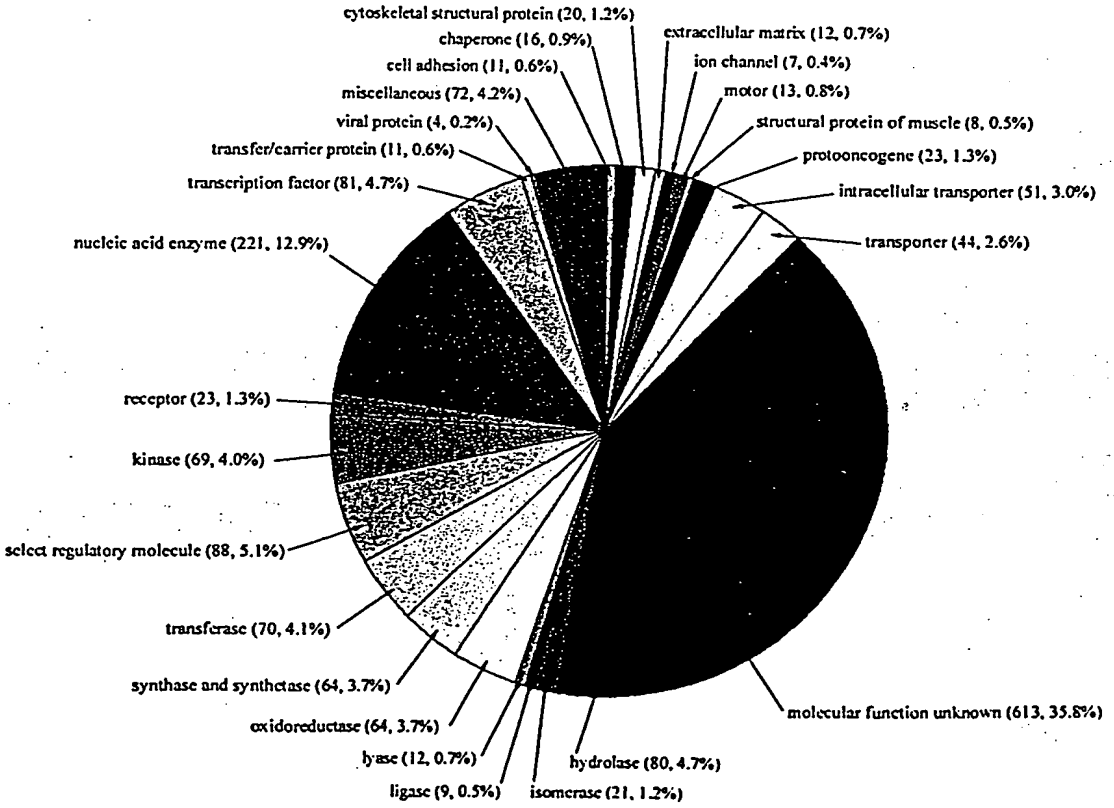


Fig. 16. Functions of putative orthologs across vertebrate and invertebrate genomes. Each slice lists the number and percentages (in parentheses) of "strict orthologs" between the human, fly, and worm genomes involved in a given category of molecular function. "Strict orthologs" are defined here as bi-directional BLAST best hits (*180*) such that each orthologous pair (i) has a BLASTP *P*-value of $\leq 10^{-10}$ (*120*), and (ii) has a more significant BLASTP score than any paralogs in either organism, i.e., there has likely been no duplication subsequent to speciation that might make the orthology ambiguous. This measure is quite strict and is a lower bound on the number of orthologs. By these criteria, there are 2758 strict human-fly orthologs, and 2031 human-worm orthologs (1523 in common between these sets).

zymes are involved in intermediary metabolism. The only exception is the hydrolase category, which is not significantly overrepresented in the shared protein set. Proteases form the largest part of this category, and several large protease families have expanded in each of these three organisms after their divergence. The category of select regulatory molecules is also overrepresented in the conserved set. The major conserved families are small guanosine triphosphatases (GTPases) (especially the Ras-related superfamily, including ADP ribosylation factor) and cell cycle regulators (particularly the cullin family, cyclin C family, and several cell division protein kinases). The last two significantly overrepresented categories are protein transport and trafficking, and chaperones. The most conserved groups in these categories are proteins involved in coated vesicle-mediated transport, and chaperones involved in protein folding and heat-shock response [particularly the DNAJ family, and heat-shock protein 60 (HSP60), HSP70, and HSP90 families]. These observations provide only a conservative estimate of the protein families in the context of specific cellular processes that were likely derived from the last common ancestor of the human, fly, and worm. As stated before, this analysis does not provide a complete estimate of conservation across the three animal genomes, as paralogous duplication makes the determination of true orthologs difficult within the members of conserved protein families.

7.3 Differences between the human genome and other sequenced eukaryotic genomes

To explore the molecular building blocks of the vertebrate taxon, we have compared the human genome with the other sequenced eukaryotic genomes at three levels: molecular functions, protein families, and protein domains.

Molecular differences can be correlated with phenotypic differences to begin to reveal the developmental and cellular processes that are unique to the vertebrates. Tables 18 and 19 display a comparison among all sequenced eukaryotic genomes, over selected protein/domain families (defined by sequence similarity, e.g., the serine-threonine protein kinases) and superfamilies (defined by shared molecular function, which may include several sequence-related families, e.g., the cytokines). In these tables we have focused on (super) families that are either very large or that differ significantly in humans compared with the other sequenced eukaryote genomes. We have found that the most prominent human expansions are in proteins involved in (i) acquired immune functions; (ii) neural development, structure, and functions; (iii) intercellular and intracellular signaling pathways in development and homeostasis; (iv) hemostasis; and (v) apoptosis.

Acquired immunity. One of the most striking differences between the human genome and the *Drosophila* or *C. elegans* genome is the appearance of genes involved in acquired immunity (Tables 18 and 19). This is expected, because the acquired immune response is a defense system that only occurs in vertebrates. We observe 22 class I and 22 class II major histocompatibility complex (MHC) antigen genes and 114 other immunoglobulin genes in the human genome. In addition, there are 59 genes in the cognate immunoglobulin receptor family. At the domain level, this is exemplified by an expansion and recruitment of the ancient immunoglobulin fold to constitute molecules such as MHC, and of the integrin fold to form several of the cell adhesion molecules that mediate interactions between immune effector cells and the extracellular matrix. Vertebrate-specific proteins include the paracrine immune regulators family of secreted 4-alpha helical bundle proteins, namely the cytokines and chemokines. Some of the cytoplasmic signal transduction components associated with cytokine receptor signal transduction are also features that are poorly represented in the fly and worm. These include protein domains found in the signal transducer and activator of transcription (STATs), the suppressors of cytokine signaling (SOCS), and protein inhibitors of activated STATs (PIAS). In contrast, many of the animal-specific protein domains that play a role in innate immune response, such as the Toll receptors, do not appear to be significantly expanded in the human genome.

Neural development, structure, and function. In the human genome, as compared with the worm and fly genomes, there is a marked increase in the number of members of protein families that are involved in neural development. Examples include neurotrophic factors such as ependymin, nerve growth factor, and signaling molecules such as semaphorins, as well as the number of proteins involved directly in neural structure and function such as myelin proteins, voltage-gated ion channels, and synaptic proteins such as synaptotagmin. These observations correlate well with the known phenotypic differences between the nervous systems of these taxa, notably (i) the increase in the number and connectivity of neurons; (ii) the increase in number of distinct neural cell types (as many as a thousand or more in human compared with a few hundred in fly and worm) (*121*); (iii) the increased length of individual axons; and (iv) the significant increase in glial cell number, especially the appearance of myelinating glial cells, which are electrically inert supporting cells differentiated from the same stem cells as neurons. A number of prominent protein expansions are involved in the processes of neural development. Of the extracellular domains that mediate cell adhesion, the connexin domain–containing proteins (*122*) exist only in humans. These proteins, which are not present in the *Drosophila* or *C. elegans* genomes, appear to provide the constitutive subunits of intercellular channels and the structural basis for electrical coupling. Pathway finding by axons and neuronal network formation is mediated through a subset of ephrins and their cognate receptor tyrosine kinases that act as positional labels to establish topographical projections (*123*). The probable biological role for the semaphorins (22 in human compared with 6 in the fly and 2 in the worm) and their receptors (neuropilins and plexins) is that of axonal guidance molecules (*124*). Signaling molecules such as neurotrophic factors and some cytokines have been shown to regulate neuronal cell survival, proliferation, and axon guidance (*125*). Notch receptors and ligands play important roles in glial cell fate determination and gliogenesis (*126*).

Other human expanded gene families play key roles directly in neural structure and function. One example is synaptotagmin (expanded more than twofold in humans relative to the invertebrates), originally found to regulate synaptic transmission by serving as a Ca^{2+} sensor (or receptor) during synaptic vesicle fusion and release (*127*). Of interest is the increased co-occurrence in humans of PDZ and the SH3 domains in neuronal-specific adaptor molecules; examples include proteins that likely modulate channel activity at synaptic junctions (*128*). We also noted expansions in several ion-channel families (Table 19), including the EAG subfamily (related to cyclic nucleotide gated channels), the voltage-gated calcium/sodium channel family, the inward-rectifier potassium channel family, and the voltage-gated potassium channel, alpha subunit family. Voltage-gated sodium and potassium channels are involved in the generation of action potentials in neurons. Together with voltage-gated calcium channels, they also play a key role in coupling action potentials to neurotransmitter release, in the development of neurites, and in short-term memory. The recent observation of a calcium-regulated association between sodium channels and synaptotagmin may have consequences for the establishment and regulation of neuronal excitability (*129*).

Myelin basic protein and myelin-associated glycoprotein are major classes of protein components in both the central and peripheral nervous system of vertebrates. Myelin P0 is a major component of peripheral myelin, and myelin proteolipid and myelin oligodendrocyte glycopotein are found in the central nervous system. Mutations in any of these

Table 18. Domain-based comparative analysis of proteins in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A). The predicted protein set of each of the above eukaryotic organisms was analyzed with Pfam version 5.5 using E value cutoffs of 0.001. The number of proteins containing the specified Pfam domains as well as the total number of domains (in parentheses) are shown in each column. Domains were categorized into cellular processes for presentation. Some domains (i.e., SH2) are listed in more than one cellular process. Results of the Pfam analysis may differ from results obtained based on human curation of protein families, owing to the limitations of large-scale automatic classifications. Representative examples of domains with reduced counts owing to the stringent E value cutoff used for this analysis are marked with a double asterisk (**). Examples include short divergent and predominantly alpha-helical domains, and certain classes of cysteine-rich zinc finger proteins.

Accession number	Domain name	Domain description	H	F	W	Y	A
		Developmental and homeostatic regulators					
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	3	0	0	0	0
PF00029	Connexin	Connexin	14 (16)	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	1	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	2	1	0	0	0
PF00007	Cys_knot	Cystine-knot domain	10 (11)	2	0	0	0
PF00778	DIX	Dix domain	5	2	4	0	0
PF00322	Endothelin	Endothelin family	3	0	0	0	0
PF00812	Ephrin	Ephrin	7 (8)	2	4	0	0
PF01404	EPh_lbd	Ephrin receptor ligand binding domain	12	2	1	0	0
PF00167	FGF	Fibroblast growth factor	23	1	1	0	0
PF01534	Frizzled	Frizzled/Smoothened family membrane region	9	7	3	0	0
PF00236	Hormone6	Glycoprotein hormones	1	0	0	0	0
PF01153	Glypican	Glypican	14	2	1	0	0
PF01271	Granin	Grainin (chromogranin or secretogranin)	3	0	0	0	0
PF02058	Guanylin	Guanylin precursor	1	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	7	4	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	10	0	0	0	0
PF02024	Leptin	Leptin	1	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	13 (23)	0	1	0	0
PF00243	NGF	Nerve growth factor family	3	0	0	0	0
PF02158	Neuregulin	Neuregulin family	4	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	1	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	3 (5)	2 (4)	2 (6)	0	0
PF00865	Osteopontin	Osteopontin	1	0	0	0	0
PF00159	Hormone3	Pancreatic hormone peptides	3	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	2	0	0	0	0
PF00123	Hormone2	Peptide hormone	5 (9)	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5	1	0	0	0
PF01403	Sema	Sema domain	27 (29)	8 (10)	3 (4)	0	0
PF01033	Somatomedin_B	Somatomedin B domain	5 (8)	3	0	0	0
PF00103	Hormone	Somatotropin	1	0	0	0	0
PF02208	Sorb	Sorbin homologous domain	2	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	3	1	1	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	17 (31)	1	0	0	0
PF00019	TGF-β	Transforming growth factor β-like domain	27 (28)	6	4	0	0
PF01099	Uteroglobin	Uteroglobin family	3	0	0	0	0
PF01160	Opiods_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	18	7 (10)	5	0	0
		Hemostasis					
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal–like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
		Immune response					
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0
PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptide	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	7	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptide	1	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	2	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
PF01582	TIR	TIR domain	18	8	2	0	131 (143)
PF00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
		PI-PY-rho GTPase signaling					
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for Ras-like GTPase	11	5	8	3	0
PF00618	RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
PF00071	Ras	Ras family	126	56 (57)	51	23	78
PF00617	RasGEF	RasGEF domain	21	8	7	5	0
PF00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
		Domains involved in apoptosis					
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0
PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
		Cytoskeletal					
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	5
PF01044	Vinculin	Vinculin family	4	2	1	0	0
		ECM adhesion					
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
		Protein interaction domains					
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	DnaJ_C	DnaJ C terminal region	12	9	5	3	19
PF00226	DnaJ	DnaJ domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17

myelin proteins result in severe demyelination, which is a pathological condition in which the myelin is lost and the nerve conduction is severely impaired (*130*). Humans have at least 10 genes belonging to four different families involved in myelin production (five myelin P0, three myelin proteolipid, myelin basic protein, and myelin-oligodendrocyte glycoprotein, or MOG), and possibly more-remotely related members of the MOG family. Flies have only a single myelin proteolipid, and worms have none at all.

Intercellular and intracellular signaling pathways in development and homeostasis. Many protein families that have expanded in humans relative to the invertebrates are involved in signaling processes, particularly in response to development and differentiation

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans Isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
		Nuclear interaction domains					
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)
PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRromatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5(6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain—N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2
PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	Zinc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

(Tables 18 and 19). They include secreted hormones and growth factors, receptors, intracellular signaling molecules, and transcription factors.

Developmental signaling molecules that are enriched in the human genome include growth factors such as wnt, transforming growth factor–β (TGF-β), fibroblast growth factor (FGF), nerve growth factor, platelet derived growth factor (PDGF), and ephrins. These growth factors affect tissue differentiation and a wide range of cellular processes involving actin-cytoskeletal and nuclear regulation. The corresponding receptors of these developmental ligands are also expanded in humans. For example, our analysis suggests at least 8 human ephrin genes (2 in the fly, 4 in the worm) and 12 ephrin receptors (2 in the fly, 1 in the worm). In the wnt signaling pathway, we find 18 wnt family genes (6 in the fly, 5 in the worm) and 12 frizzled receptors (6 in the fly, 5 in the worm). The Groucho family of transcriptional corepressors downstream in the wnt pathway are even more markedly expanded, with 13 predicted members in humans (2 in the fly, 1 in the worm).

Extracellular adhesion molecules involved in signaling are expanded in the human genome (Tables 18 and 19). The interactions of several of these adhesion domains with extracellular matrix proteoglycans play a critical role in host defense, morphogenesis, and tissue repair (*131*). Consistent with the well-defined role of heparan sulfate proteoglycans in modulating these interactions (*132*), we observe an expansion of the heparin sulfate sulfotransferases in the human genome relative to worm and fly. These sulfotransferases modulate tissue differentiation (*133*). A similar expansion in humans is noted in structural proteins that constitute the actin-cytoskeletal architecture. Compared with the fly and worm, we observe an explosive expansion of the nebulin (35 domains per protein on average), aggrecan (12 domains per protein on average), and plectin (5 domains per protein on average) repeats in humans. These repeats are present in proteins involved in modulating the actin-cytoskeleton with predominant expression in neuronal, muscle, and vascular tissues.

Comparison across the five sequenced eukaryotic organisms revealed several expanded protein families and domains involved in cytoplasmic signal transduction (Table 18). In particular, signal transduction pathways playing roles in developmental regulation and acquired immunity were substantially enriched. There is a factor of 2 or greater expansion in humans in the Ras superfamily GTPases and the GTPase activator and GTP exchange factors associated with them. Although there are about the same number of tyrosine kinases in the human and *C. elegans* genomes, in humans there is an increase in the SH2, PTB, and ITAM domains involved in phosphotyrosine signal transduction. Further, there is a twofold expansion of phosphodiesterases in the human genome compared with either the worm or fly genomes.

The downstream effectors of the intracellular signaling molecules include the transcription factors that transduce developmental fates. Significant expansions are noted in the ligand-binding nuclear hormone receptor class of transcription factors compared with the fly genome, although not to the extent observed in the worm (Tables 18 and 19). Perhaps the most striking expansion in humans is in the C2H2 zinc finger transcription factors. Pfam detects a total of 4500 C2H2 zinc finger domains in 564 human proteins, compared with 771 in 234 fly proteins. This means that there has been a dramatic expansion not only in the number of C2H2 transcription factors, but also in the number of these DNA-binding motifs per transcription factor (8 on average in humans, 3.3 on average in the fly, and 2.3 on average in the worm). Furthermore, many of these transcription factors contain either the KRAB or SCAN domains, which are not found in the fly or worm genomes. These domains are involved in the oligomerization of transcription factors and increase the combinatorial partnering of these factors. In general, most of the transcription factor domains are shared between the three animal genomes, but the reassortment of these domains results in organism-specific transcription factor families. The domain combinations found in the human, fly, and worm include the BTB with C2H2 in the fly and humans, and homeodomains alone or in combination with Pou and LIM domains in all of the animal genomes. In plants, however, a different set of transcription factors are expanded, namely, the myb family, and a unique set that includes VP1 and AP2 domain–containing proteins (*134*). The yeast genome has a paucity of transcription factors compared with the multicellular eukaryotes, and its repertoire is limited to the expansion of the yeast-specific C6 transcription factor family involved in metabolic regulation.

While we have illustrated expansions in a subset of signal transduction molecules in the human genome compared with the other eukaryotic genomes, it should be noted that most of the protein domains are highly conserved. An interesting observation is that worms and humans have approximately the same number of both tyrosine kinases and serine/threonine kinases (Table 19). It is important to note, however, that these are merely counts of the catalytic domain; the proteins that contain these domains also display a wide repertoire of interaction domains with significant combinatorial diversity.

Hemostasis. Hemostasis is regulated primarily by plasma proteases of the coagulation pathway and by the interactions that occur between the vascular endothelium and platelets. Consistent with known anatomical and physiological differences between vertebrates and invertebrates, extracellular adhesion domains that constitute proteins integral to hemostasis are expanded in the human relative to the fly and worm (Tables 18 and 19). We note the evolution of domains such as FIMAC, FN1, FN2, and C1q that mediate surface interactions between hematopoeitic cells and the vascular matrix. In addition, there has been extensive recruitment of more-ancient animal-specific domains such as VWA, VWC, VWD, kringle, and FN3 into multidomain proteins that are involved in hemostatic regulation. Although we do not find a large expansion in the total number of serine proteases, this enzymatic domain has been specifically recruited into several of these multidomain proteins for proteolytic regulation in the vascular compartment. These are represented in plasma proteins that belong to the kinin and complement pathways. There is a

significant expansion in two families of matrix metalloproteases: ADAM (a disintegrin and metalloprotease) and MMPs (matrix metalloproteases) (Table 19). Proteolysis of extracellular matrix (ECM) proteins is critical for tissue development and for tissue degradation in diseases such as cancer, arthritis, Alzheimer's disease, and a variety of inflammatory conditions (*135*, *136*). ADAMs are a family of integral membrane proteins with a pivotal role in fibrinogenolysis and modulating interactions between hematopoietic components and the vascular matrix components. These proteins have been shown to cleave matrix proteins, and even signaling molecules: ADAM-17 converts tumor necrosis factor–α, and ADAM-10 has been implicated in the Notch signaling pathway (*135*). We have identified 19 members of the matrix metalloprotease family, and a total of 51 members of the ADAM and ADAM-TS families.

Apoptosis. Evolutionary conservation of some of the apoptotic pathway components across eukarya is consistent with its central role in developmental regulation and as a response to pathogens and stress signals. The signal transduction pathways involved in programmed cell death, or apoptosis, are mediated by interactions between well-characterized domains that include extracellular domains, adaptor (protein-protein interaction) domains, and those found in effector and regulatory enzymes (*137*). We enumerated the protein counts of central adaptor and effector enzyme domains that are found only in the apoptotic pathways to provide an estimate of divergence across eukarya and relative expansion in the human genome when compared with the fly and worm (Table 18). Adaptor domains found in proteins restricted only to apoptotic regulation such as the DED domains are vertebrate-specific, whereas others like BIR, CARD, and Bcl2 are represented in the fly and worm (although the number of Bcl2 family members in humans is significantly expanded). Although plants and yeast lack the caspases, caspase-like molecules, namely the para- and meta-caspases, have been reported in these organisms (*138*). Compared with other animal genomes, the human genome shows an expansion in the adaptor and effector domain–containing proteins involved in apoptosis, as well as in the proteases involved in the cascade such as the caspase and calpain families.

Expansions of other protein families. *Metabolic enzymes.* There are fewer cytochrome P450 genes in humans than in either the fly or worm. Lipoxygenases (six in humans), on the other hand, appear to be specific to the vertebrates and plants, whereas the lipoxygenase-activating proteins (four in humans) may be vertebrate-specific. Lipoxygenases are involved in arachidonic acid metabolism, and they and their activators have been implicated in diverse human pathology ranging from allergic responses to cancers. One of the most surprising human expansions, however, is in the number of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (46 in humans, 3 in the fly, and 4 in the worm). There is, however, evidence for many retrotransposed GAPDH pseudogenes (*139*), which may account for this apparent expansion. However, it is interesting that GAPDH, long known as a conserved enzyme involved in basic metabolism found across all phyla from bacteria to humans, has recently been shown to have other functions. It has a second cat-

Table 19. Number of proteins assigned to selected Panther families or subfamilies in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A).

Panther family/subfamily*	H	F	W	Y	A
Neural structure, function, development					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated Ca^{2+} alpha	22	4	8	2	2
Voltage-gated Ca^{2+} alpha-2	10	3	2	0	0
Voltage-gated Ca^{2+} beta	5	2	2	0	0
Voltage-gated Ca^{2+} gamma	1	0	0	0	0
Voltage-gated K^+ alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated Na^+	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
Immune response					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Inteferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small inducible cytokine A	14	0	0	0	0
SI cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor†	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
FI cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor†	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-immunoglobulin receptor	4	0	0	0	0

alytic activity, as a uracil DNA glycosylase (*140*) and functions as a cell cycle regulator (*141*) and has even been implicated in apoptosis (*142*).

Translation. Another striking set of human expansions has occurred in certain families involved in the translational machinery. We identified 28 different ribosomal subunits that each have at least 10 copies in the genome; on average, for all ribosomal proteins there is about an 8- to 10-fold expansion in the number of genes relative to either the worm or fly. Retrotransposed pseudogenes may account for many of these expansions [see the discussion above and (*143*)]. Recent evidence suggests that a number of ribosomal proteins have secondary functions independent of their involvement in protein biosynthesis; for example, L13a and the related L7 subunits (36 copies in humans) have been shown to induce apoptosis (*144*).

There is also a four- to fivefold expansion in the elongation factor 1-alpha family (eEF1A; 56 human genes). Many of these expansions likely represent intronless paralogs that have presumably arisen from retrotransposition, and again there is evidence that many of these may be pseudogenes (*145*). However, a second form (eEF1A2) of this factor has been identied with tissue-specific expression in skeletal muscle and a complementary expression pattern to the ubiquitously expressed eEF1A (*146*).

Ribonucleoproteins. Alternative splicing results in multiple transcripts from a single gene, and can therefore generate additional diversity in an organism's protein complement. We have identified 269 genes for ribonucleoproteins. This represents over 2.5 times the number of ribonucleoprotein genes in the worm, two times that of the fly, and about the same as the 265 identified in the *Arabidopsis* genome. Whether the diversity of ribonucleoprotein genes in humans contributes to gene regulation at either the splicing or translational level is unknown.

Posttranslational modifications. In this set of processes, the most prominent expansion is the transglutaminases, calcium-dependent enzymes that catalyze the cross-linking of proteins in cellular processes such as hemostasis and apoptosis (*147*). The vitamin K–dependent gamma carboxylase gene product acts on the GLA domain (missing in the fly and worm) found in coagulation factors, osteocalcin, and matrix GLA protein (*148*). Tyrosylprotein sulfotransferases participate in the posttranslational modification of proteins involved in inflammation and hemostasis, including coagulation factors and chemokine receptors (*149*). Although there is no significant numerical increase in the counts for domains involved in nuclear protein modification, there are a number of domain arrangements in the predicted human proteins that are not found in the other currently sequenced genomes. These include the tandem association of two histone deacetylase domains in HD6 with a ubiquitin finger domain, a feature lacking in the fly genome. An additional example is the co-occurrence of important nuclear regulatory enzyme PARP (poly-ADP ribosyl transferase) domain fused to protein-interaction domains—BRCT and VWA in humans.

Concluding remarks. There are several possible explanations for the differences in phenotypic complexity observed in humans when compared to the fly and worm. Some of these relate to the prominent differences in the immune system, hemostasis, neuronal, vascular, and cytoskeletal complexity. The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the levels of protein architecture, transcriptional and translational control, posttranslational modification of proteins, or posttranscriptional regulation. Extensive domain shuffling to increase or alter combinatorial diversity can provide an exponential

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other immunoglobulin†	114	0	0	0	0
Toll receptor–related	10	6	0	0	0
Developmental and homeostatic regulators					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0
Neuregulin/heregulin	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoeitin	2	0	1	0	0
Thyomosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
Kinases and phosphatases					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
Signal transduction					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors†‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0

increase in the ability to mediate protein-protein interactions without dramatically increasing the absolute size of the protein complement (*150*). Evolution of apparently new (from the perspective of sequence analysis) protein domains and increasing regulatory complexity by domain accretion both quantitatively and qualitatively (recruitment of novel domains with preexisting ones) are two features that we observe in humans. Perhaps the best illustration of this trend is the C2H2 zinc finger–containing transcription factors, where we see expansion in the number of domains per protein, together with vertebrate-specific domains such as KRAB and SCAN. Recent reports on the prominent use of internal ribosomal entry sites in the human genome to regulate translation of specific classes of proteins suggests that this is an area that needs further research to identify the full extent of this process in the human genome (*151*). At the posttranslational level, although we provide examples of expansions of some protein families involved in these modifications, further experimental evidence is required to evaluate whether this is correlated with increased complexity in protein processing. Posttranscriptional processing and the extent of isoform generation in the human remain to be cataloged in their entirety. Given the conserved nature of the spliceosomal machinery, further analysis will be required to dissect regulation at this level.

8 Conclusions

8.1 The whole-genome sequencing approach versus BAC by BAC

Experience in applying the whole-genome shotgun sequencing approach to a diverse group of organisms with a wide range of genome sizes and repeat content allows us to assess its strengths and weaknesses. With the success of the method for a large number of microbial genomes, *Drosophila*, and now the human, there can be no doubt concerning the utility of this method. The large number of microbial genomes that have been sequenced by this method (*15*, *80*, *152*) demonstrate that megabase-sized genomes can be sequenced efficiently without any input other that the de novo mate-paired sequences. With more complex genomes like those of *Drosophila* or human, map information, in the form of well-ordered markers, has been critical for long-range ordering of scaffolds. For joining scaffolds into chromosomes, the quality of the map (in terms of the order of the markers) is more important than the number of markers per se. Although this mapping could have been performed concurrently with sequencing, the prior existence of mapping data was beneficial. During the sequencing of the *A. thaliana* genome, sequencing of individual BAC clones permitted extension of the se-

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
Transcription factors/chromatin organization					
C2H2 zinc finger–containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0
Forkhead-related	34	19	15	4	0
FOS	8	2	1	0	0
Groucho	13	2	1	0	0
Histone H1	5	0	1	0	0
Histone H2A	24	1	17	3	13
Histone H2B	21	1	17	2	12
Histone H3	28	2	24	2	16
Histone H4	9	1	16	1	8
Homeotic†	168	104	74	4	78
ABD-B	5	0	0	0	0
Bithoraxoid	1	8	1	0	0
Iroquois class	7	3	1	0	0
Distal-less	5	2	1	0	0
Engrailed	2	2	1	0	0
LIM-containing	17	8	3	0	0
MEIS/KNOX class	9	4	4	2	26
NK-3/NK-2 class	9	4	5	0	0
Paired box	38	28	23	0	2
Six	5	3	4	0	0
Leucine zipper	6	0	0	0	0
Nuclear hormone receptor†	59	25	183	1	4
Pou-related	15	5	4	1	0
Runt-related	3	4	2	0	0
ECM adhesion					
Cadherin	113	17	16	0	0
Claudin	20	0	0	0	0
Complement receptor-related	22	8	6	0	0
Connexin	14	0	0	0	0
Galectin	12	5	22	0	0
Glypican	13	2	1	0	0
ICAM	6	0	0	0	0
Integrin alpha	24	7	4	0	1
Integrin beta	9	2	2	0	0
LDL receptor family	26	19	20	0	2
Proteoglycans	22	9	7	0	5
Apoptosis					
Bcl-2	12	1	0	0	0
Calpain	22	4	11	1	3
Calpain inhibitor	4	0	0	0	1
Caspase	13	7	3	0	0
Hemostasis					
ADAM/ADAMTS	51	9	12	0	0
Fibronectin	3	0	0	0	0
Globin	10	2	3	0	3
Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
Other enzymes					
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
Splicing and translation					
EF-1alpha	56	13	10	6	13
Ribonucleoproteins†	269	135	104	60	265
Ribosomal proteins†	812	111	80	117	256

*The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models. †This class represents a number of different families in the same Panther molecular function subcategory. ‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

quence well into centromeric regions and allowed high-quality resolution of complex repeat regions. Likewise, in *Drosophila*, the BAC physical map was most useful in regions near the highly repetitive centromeres and telomeres. WGA has been found to deliver excellent-quality reconstructions of the unique regions of the genome. As the genome size, and more importantly the repetitive content, increases, the WGA approach delivers less of the repetitive sequence.

The cost and overall efficiency of clone-by-clone approaches makes them difficult to justify as a stand-alone strategy for future large-scale genome-sequencing projects. Specific applications of BAC-based or other clone mapping and sequencing strategies to resolve ambiguities in sequence assembly that cannot be efficiently resolved with computational approaches alone are clearly worth exploring. Hybrid approaches to whole-genome sequencing will only work if there is sufficient coverage in both the whole-genome shotgun phase and the BAC clone sequencing phase. Our experience with human genome assembly suggests that this will require at least 3× coverage of both whole-genome and BAC shotgun sequence data.

8.2 The low gene number in humans

We have sequenced and assembled ~95% of the euchromatic sequence of *H. sapiens* and used a new automated gene prediction method to produce a preliminary catalog of the human genes. This has provided a major surprise: We have found far fewer genes (26,000 to 38,000) than the earlier molecular predictions (50,000 to over 140,000). Whatever the reasons for this current disparity, only detailed annotation, comparative genomics (particularly using the *Mus musculus* genome), and careful molecular dissection of complex phenotypes will clarify this critical issue of the basic "parts list" of our genome. Certainly, the analysis is still incomplete and considerable refinement will occur in the years to come as the precise structure of each transcription unit is evaluated. A good place to start is to determine why the gene estimates derived from EST data are so discordant with our predictions. It is likely that the following contribute to an inflated gene number derived from ESTs: the variable lengths of 3′- and 5′-untranslated leaders and trailers; the little-understood vagaries of RNA processing that often leave intronic regions in an unspliced condition; the finding that nearly 40% of human genes are alternatively spliced (*153*); and finally, the unsolved technical problems in EST library construction where contamination from heterogeneous nuclear RNA and genomic DNA are not uncommon. Of course, it is possible that there are genes that remain unpredicted owing to the absence of EST or protein data to support them, although our use of mouse genome data for predicting genes should limit this number. As was true at the beginning of genome sequencing, ultimately it will be necessary to measure mRNA in specific cell types to demonstrate the presence of a gene.

J. B. S. Haldane speculated in 1937 that a population of organisms might have to pay a price for the number of genes it can possibly carry. He theorized that when the number of genes becomes too large, each zygote carries so many new deleterious mutations that the population simply cannot maintain itself. On the basis of this premise, and on the basis of available mutation rates and x-ray–induced mutations at specific loci, Muller, in 1967 (*154*), calculated that the mammalian genome would contain a maximum of not much more than 30,000 genes (*155*). An estimate of 30,000 gene loci for humans was also arrived at by Crow and Kimura (*156*). Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived by annotation of the fly genome (*26*, *27*). These arguments for the theoretical maximum gene number were based on simplified ideas of genetic load—that all genes have a certain low rate of mutation to a deleterious state. However, it is clear that many mouse, fly, worm, and yeast knockout mutations lead to almost no discernible phenotypic perturbations.

The modest number of human genes means that we must look elsewhere for the mechanisms that generate the complexities inherent in human development and the sophisticated signaling systems that maintain homeostasis. There are a large number of ways in which the functions of individual genes and gene products are regulated. The degree of "openness" of chromatin structure and hence transcriptional activity is regulated by protein complexes that involve histone and DNA enzymatic modifications. We enumerate many of the proteins that are likely involved in nuclear regulation in Table 19. The location, timing, and quantity of transcription are intimately linked to nuclear signal transduction events as well as by the tissue-specific expression of many of these proteins. Equally important are regulatory DNA elements that include insulators, repeats, and endogenous viruses (*157*); methylation of CpG islands in imprinting (*158*); and promoter-enhancer and intronic regions that modulate transcription. The spliceosomal machinery consists of multisubunit proteins (Table 19) as well as structural and catalytic RNA elements (*159*) that regulate transcript structure through alternative start and termination sites and splicing. Hence, there is a need to study different classes of RNA molecules (*160*) such as small nucleolar RNAs, antisense riboregulator RNA, RNA involved in X-dosage compensation, and other structural RNAs to appreciate their precise role in regulating gene expression. The phenomenon of RNA editing in which coding changes occur directly at the level of mRNA is of clinical and biological relevance (*161*). Finally, examples of translational control include internal ribosomal entry sites that are found in proteins involved in cell cycle regulation and apoptosis (*162*). At the protein level, minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology (*163*). This dynamic system therefore has many ways to modulate activity, which suggests that definition of complex systems by analysis of single genes is unlikely to be entirely successful.

In situ studies have shown that the human genome is asymmetrically populated with G+C content, CpG islands, and genes (*68*). However, the genes are not distributed quite as unequally as had been predicted (Table 9) (*69*). The most G+C-rich fraction of the genome, H3 isochores, constitute more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C L isochores make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the "desertification" of the vertebrate genome (*71*). Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution? If these deserts are dispensable, it ought to be possible to find mammalian genomes that are far smaller in size than the human genome. Indeed, many species of bats have genome sizes that are much smaller than that of humans; for example, *Miniopterus*, a species of Italian bat, has a genome size that is only 50% that of humans (*164*). Similarly, *Muntiacus*, a species of Asian barking deer, has a genome size that is ~70% that of humans.

8.3 Human DNA sequence variation and its distribution across the genome

This is the first eukaryotic genome in which a nearly uniform ascertainment of polymorphism has been completed. Although we have identified and mapped more than 3 million SNPs, this by no means implies that the task of finding and cataloging SNPs is complete. These represent only a fraction of the SNPs present in the human population as a whole. Nevertheless, this first glimpse at genome-wide variation has revealed strong inhomogeneities in the distribution of SNPs across the genome. Polymorphism in DNA carries with it a snapshot of the past operation of population genetic forces, including mutation, migration, selection, and genetic drift. The availability of a dense array of SNPs will allow questions related to each of these factors to be addressed on a genome-wide basis. SNP studies can establish the range of haplo-

types present in subjects of different ethnogeographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of interspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele "swept" the linked variation out of the population (*165*). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (*166*). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of "this binds to that, which then docks on this, and then the complex moves there. . . ." (*167*) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other "parts lists" reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (*168*). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (*169*), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about 1.5 cm^3, two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are "special cases" of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of "regulatory genes" that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF-β, ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these "complexities" perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein's brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (*170*). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (*171*). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (*172*), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no "good" genes or "bad" genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with "complexity," particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human ge-

nome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

References and Notes

1. R. L. Sinsheimer, *Genomics* 5, 954 (1989); U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
2. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
3. F. Sanger *et al.*, *Nature* 265, 687 (1977).
4. P. H. Seeburg *et al.*, *Trans. Assoc. Am. Physicians* 90, 109 (1977).
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* 154, 353 (1986).
6. J. Gocayne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 84, 8296 (1987).
7. A. Martin-Gallardo *et al.*, *DNA Sequence* 3, 237 (1992); W. R. McCombie *et al.*, *Nature Genet.* 1, 348 (1992); M. A. Jensen *et al.*, *DNA Sequence* 1, 233 (1991).
8. M. D. Adams *et al.*, *Science* 252, 1651 (1991).
9. M. D. Adams *et al.*, *Nature* 355, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* 4, 373 (1993); M. H. Polymeropoulos *et al.*, *Nature Genet.* 4, 381 (1993); M. Marra *et al.*, *Nature Genet.* 21, 191 (1999).
10. M. D. Adams *et al.*, *Nature* 377, 3 (1995); O. White *et al.*, *Nucleic Acids Res.* 21, 3829 (1993).
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* 162, 729 (1982).
12. B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* 57, 577 (1991).
13. R. D. Fleischmann *et al.*, *Science* 269, 496 (1995).
14. C. M. Fraser *et al.*, *Science* 270, 397 (1995).
15. C. J. Bult *et al.*, *Science* 273, 1058 (1996); J. F. Tomb *et al.*, *Nature* 388, 539 (1997); H. P. Klenk *et al.*, *Nature* 390, 364 (1997).
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* 381, 364 (1996).
17. H. Schmitt *et al.*, *Genomics* 33, 9 (1996).
18. S. Zhao *et al.*, *Genomics* 63, 321 (2000).
19. X. Lin *et al.*, *Nature* 402, 761 (1999).
20. J. L. Weber, E. W. Myers, *Genome Res.* 7, 401 (1997).
21. P. Green, *Genome Res.* 7, 410 (1997).
22. E. Pennisi, *Science* 280, 1185 (1998).
23. J. C. Venter *et al.*, *Science* 280, 1540 (1998).
24. M. D. Adams *et al.*, *Nature* 368, 474 (1994).
25. E. Marshall, E. Pennisi, *Science* 280, 994 (1998).
26. M. D. Adams *et al.*, *Science* 287, 2185 (2000).
27. G. M. Rubin *et al.*, *Science* 287, 2204 (2000).
28. E. W. Myers *et al.*, *Science* 287, 2196 (2000).
29. F. S. Collins *et al.*, *Science* 282, 682 (1998).
30. International Human Genome Sequencing Consortium (2001), *Nature* 409, 860 (2001).
31. Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
32. Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotype analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.
33. DNA was isolated from blood (*173*) or sperm. For sperm, a washed pellet (100 μl) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl–20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3′-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3′-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with Bgl II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp Bam HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of Bgl II. As Bgl II–Bgl II ligations occurred, they were continually cleaved, whereas Bam HI–Bgl II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50 μg/ml), carbenicillin (50 μg/ml), and kanamycin (15 μg/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.
34. Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (*173*) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a

sample plate barcode, thus enhancing sample sheet-to-plate associations.
35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977); J. M. Prober *et al.*, *Science* **238**, 336 (1987).
36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).
37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (*174*), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5′ vector splice junction in the initial part of a read was removed.
38. National Center for Biotechnology Information (NCBI); available at www.ncbi.nlm.nih.gov/.
39. NCBI; available at www.ncbi.nlm.nih.gov/HTGS/.
40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.
41. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (*175*), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.
42. G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* **3**, 563 (1996).
43. E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73–89.
44. P. Deloukas *et al.*, *Science* **282**, 744 (1998).
45. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).
46. J. Zhang *et al.*, data not shown.
47. Shredded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.
48. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
49. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000).
51. The International RH Mapping Consortium, available at www.ncbi.nlm.nih.gov/genemap99/.
52. See http://ftp.genome.washington.edu/RM/RepeatMasker.html.
53. G. D. Schuler, *Trends Biotechnol.* **16**, 456 (1998).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
55a. M. Olivier *et al.*, *Science* **291**, 1298 (2001).
55b. See http://genome.ucsc.edu/.
56. N. Chaudhari, W. E. Hahn, *Science* **220**, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983).
57. D. Dickson, *Nature* **401**, 311 (1999).
58. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
59. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
60. M. Yandell, in preparation.
61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000).
62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (*52*) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3×), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (*54*) optimized for the Compaq Alpha computeserver and an effective database size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, human nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and rodent nucleotide BLAST results having an expectation score of $<1 \times 10^{9}$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (*63*).
63. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996); C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); R. J. Mural, *Methods Enzymol.* **303**, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); Floreal *et al.*, *Genome Res.* **8**, 967 (1998).
64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* **31**, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994).
65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.
66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000).
67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* **5**, 144 (1989).
68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).
69. G. Bernardi, *Gene* **241**, 3 (2000).
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).
71. S. Ohno, *Trends Genet.* **1**, 160 (1985).
72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* **63**, 861 (1998).
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000).
74. A. Bird, *Trends Genet.* **3**, 342 (1987).
75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).
77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995).
78. J. Peters, *Genome Biol.* **1**, reviews1028.1 (2000) (http://genomebiology.com/2000/1/5/reviews/1028).
79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* **9**, 2651 (2000).
80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993).
81. S. H. Cross *et al.*, *Mamm. Genome* **11**, 373 (2000).
82. D. Slavov *et al.*, *Gene* **247**, 215 (2000).
83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).
84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000).
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999).
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999).
87. P. Nouvel, *Genetica* **93**, 191 (1994).
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000).
89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair *ij* in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by *i* and *j*. This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B–containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they

share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (*27*) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: If one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

90. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
91. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (1999).
92. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A′ and B′ (paralogs of A and B) also exist in the genome, the probability that B′ occurs immediately after A′ is $1/N$, where N is the number of proteins in the set (for this analysis, N = 26,588). Allowing for B′ to occur as any of the next J-1 proteins [leaving a gap between A′ and B′ increases the probability to $(J - 1)/N$; allowing B′A′ or A′B′ gives a probability of $2(J - 1)/N$]. Considering three genes ABC, the probability of observing A′B′C′ elsewhere in the genome, given that the paralogs exist, is $1/N^2$. Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that K proteins can be spread across J positions by counting all possible arrangements of $K - 2$ proteins in the $J - 2$ positions between the first and last protein. Allowing for a spread to vary from K positions (no gaps) to J gives
$$L = \sum_{X=K-2}^{J-2} \binom{X}{K-2}$$
arrangements. Thus, the probability of chance occurrence is L/N^{K-1}. Allowing for both sets of genes (e.g., ABC and A′B′C′) to be spread across J positions increases this to L^2/N^{K-1}. The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for M such rearrangements gives us a probability $P = L^2M/N^{K-1}$. For example, the probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is $36/N^2$; the expected number of such matched sets in the predicted protein set is approximately $(N)36/N^2 = 36/N$, a value $\ll 1$. Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with $P \ll 1$.
94. B. J. Trask *et al.*, *Hum. Mol. Genet.* **7**, 13 (1998); D. Sharon *et al.*, *Genomics* **61**, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* **10**, 1351 (2000); A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000); D. W. Burt *et al.*, *Nature* **402**, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998); P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999).
98. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
99. G. T. Marth *et al.*, *Nature Genet.* **23**, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* **7**, 649 (1997).
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage x from a given individual, both homologs are present in the assembly with probability $1 - (1/2)x^{-1}$. Even if both homologs are present, the probability that a SNP is detected is <1 because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.
106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* **150**, 1133 (1998).
107. D. A. Nickerson *et al.*, *Nature Genet.* **19**, 233 (1998); D. A. Nickerson *et al.*, *Genomic Res.* **10**, 1532 (2000); L. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000); D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
109. S. Tavare, *Theor. Popul. Biol.* **26**, 119 (1984).
110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1–44.
111. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* **22**, 78 (1999).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value $< 10^{-5}$) and "globally" alignable (the length of the match region must be >70% and <130% of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2. The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NLL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (*176*) were then used as input into the BETE program (*177*). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (*178*) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NLL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.
117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* **27**, 229 (1999).
118. A. Goffeau *et al.*, *Science* **274**, 546, 563 (1996).
119. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
120. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
121. E. R. Kandel, J. H. Schwartz, T. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).
122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* **65**, 475 (1996).
123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* **407**, 963 (2000); P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* **27**, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).
130. G. Lemke, *Glia* **7**, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* **404**, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998).
134. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* **274**, 25555 (1999).

136. R. A. Black, J. M. White, *Curr. Opin. Cell Biol.* **10**, 654 (1998).
137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999).
138. A. G. Uren *et al.*, *Mol. Cell* **6**, 961 (2000).
139. P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* **4**, 695 (1993).
140. K. Meyer-Siegler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8460 (1991).
141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* **21**, 993 (1993).
142. N. A. Tatton, *Exp. Neurol.* **166**, 29 (2000).
143. N. Kenmochi *et al.*, *Genome Res.* **8**, 509 (1998).
144. F. W. Chen, Y. A. Ioannou, *Int. Rev. Immunol.* **18**, 429 (1999).
145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* **18**, 1513 (1990).
146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4463 (1998); A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* **216**, 267 (1999).
147. D. Aeschlimann, V. Thomazy, *Connect. Tissue Res.* **41**, 1 (2000).
148. P. Munroe *et al.*, *Nature Genet.* **21**, 142 (1999); S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* **254**, 1634 (1991); B. Furie *et al.*, *Blood* **93**, 1798 (1999).
149. J. W. Kehoe, C. R. Bertozzi, *Chem. Biol.* **7**, R57 (2000).
150. T. Pawson, P. Nash, *Genes Dev.* **14**, 1027 (2000).
151. A. W. van der Velden, A. A. Thomas, *Int. J. Biochem. Cell Biol.* **31**, 87 (1999).
152. C. M. Fraser *et al.*, *Science* **281**, 375 (1998); H. Tettelin *et al.*, *Science* **287**, 1809 (2000).
153. D. Brett *et al.*, *FEBS Lett.* **474**, 83 (2000).
154. H. J. Muller, H. Kern, *Z. Naturforsch. B* **22**, 1330 (1967).
155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.
156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
157. K. Kobayashi *et al.*, *Nature* **394**, 388 (1998).
158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* **249**, 87 (2000).
159. C. A. Collins, C. Guthrie, *Nature Struct. Biol.* **7**, 850 (2000).
160. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999).
161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000).
162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* **16**, 469 (2000).
163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* **408**, 106 (2000).
164. E. Capanna, M. G. M. Romanini, *Caryologia* **24**, 471 (1971).
165. J. Maynard Smith, *J. Theor. Biol.* **128**, 247 (1987).
166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995).
167. J. E. Bailey, *Nature Biotechnol.* **17**, 616 (1999).
168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3731 (1998).
169. G. L. Miklos, *J. Neurobiol.* **24**, 842 (1993).
170. J. P. Crutchfield, K. Young, *Phys. Rev. Lett.* **63**, 105 (1989); M. Gell-Mann, S. Lloyd, *Complexity* **2**, 44 (1996).
171. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
172. E. Colucci-Guyon *et al.*, *Cell* **79**, 679 (1994).
173. J. Sambrook, E. F. Fritch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).
174. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).
175. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
176. A. Krogh, K. Sjölander, *J. Mol. Biol.* **235**, 1501 (1994).
177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* **6**, 165 (1998).
178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000).
179. GO, available at www.geneontology.org/.
180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000).

181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site (www.celera.com). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001

EXHIBIT B

THE HUMAN GENOME

Humanity has been given a great gift. With the completion of the human genome sequence, we have received a powerful tool for unlocking the secrets of our genetic heritage and for finding our place among the other participants in the adventure of life.

This week's issue of *Science* contains the report of the sequencing of the human genome from a group of authors led by Craig Venter of Celera Genomics. The report of the sequencing of the human genome from the publicly funded consortium of laboratories led by Francis Collins appears in this week's *Nature*. This stunning achievement has been portrayed—often unfairly—as a competition between two ventures, one public and one private. That characterization detracts from the awesome accomplishment jointly unveiled this week. In truth, each project contributed to the other. The inspired vision that launched the publicly funded project roughly 10 years ago reflected, and now rewards, the confidence of those who believe that the pursuit of large-scale fundamental problems in the life sciences is in the national interest. The technical innovation and drive of Craig Venter and his colleagues made it possible to celebrate this accomplishment far sooner than was believed possible. Thus, we can salute what has become, in the end, not a contest but a marriage (perhaps encouraged by shotgun) between public funding and private entrepreneurship.

A historic moment for the scientific endeavor.

There are excellent scientific reasons for applauding an outcome that has given us two winners. Two sequences are better than one; the opportunity for comparison and convergence is invaluable. Indeed, a real-world proof of the importance of access to both sets of data can be found in the pages of this issue of *Science*, in the comparative analysis by Olivier *et al.* (p. 1298).

Although we have made the point before, it is worth repeating that the sequencing of the human genome represents, not an ending, but the beginning of a new approach to biology. As Galas says in his Viewpoint (p. 1257), the knowledge that all of the genetic components of any process can be identified will give extraordinary new power to scientists. Because of this breakthrough, research can evolve from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. Several articles in this issue highlight how this approach is already beginning to revolutionize the way we look at human disease.

This has been a massive project, on a scale unparalleled in the history of biology, but of course it has built on the scientific insights of centuries of investigators. By coincidence, this landmark announcement falls during the week of the anniversary of the birth of Charles Darwin. Darwin's message that the survival of a species can depend on its ability to evolve in the face of change is peculiarly pertinent to discussions that have gone on in the past year over access to the Celera data. (Full information regarding the agreements that were reached to make the data available can be found at www.sciencemag.org/feature/data/announcement/gsp.shl.) We are willing to be flexible in allowing data repositories other than the traditional GenBank, while insisting on access to all the data needed to verify conclusions. In this domain, change is everywhere: Commercial researchers are producing more and more potentially valuable sequences, yet (at least in the United States) laws governing databases provide scant protection against piracy. Had the Celera data been kept secret, it would have been a serious loss to the scientific community. We hope that our adaptability in the face of change will enable other proprietary data to be published after peer review, in a way that satisfies our continuing commitment to full access.

It should be no surprise that an achievement so stunning, and so carefully watched, has created new challenges for the scientific venture. *Science* is proud to have played a role in bringing this discovery onto the public stage. It is literally true that this is a historic moment for the scientific endeavor. The human genome has been called the Book of Life. Rather, it is a library, in which, with rules that encourage exploration and reward creativity, we can find many of the books that will help define us and our place in the great tapestry of life.

Barbara R. Jasny and Donald Kennedy

How to contact us

If you are unable to obtain the information you need from our website, please contact us for more information about services, as well as sample preparation guidelines.

Web Page: http://info.med.yale.edu/wmkeck/dna_arrays.htm

Campus Location: Temp. location: SHMC-E16 - Sterling Hall of Medicine C-wing. Perm. Location WWW 5 - Basement Level Yale Cancer Center, Winchester Wing

Campus Mailing Address: YCC/HHMI Biopolymer/Keck Biotechnology Resource Laboratory, Array Technology Section, 333 Cedar Street, WWW 5, New Haven, CT 06520

Co-Director: Kenneth Williams, Ph.D.

Telephone: 203-737-2206

FAX: 203-737-2638

Email: kenneth.williams@yale.edu

Co-Director: Archibald Perkins, M.D., Ph.D.

Telephone: 203-785-6843

FAX: 203-785-7467

Email: archibald.perkins@yale.edu

Manager: Janet Hager, Ph.D.

Telephone: 203-785-7945

FAX: 203-785-5259

Email: janet.hager@yale.edu

Array Lab: SHMC-E16
Telephone: 203-785-4950
Email: keck.array@yale.edu

Home

Array Technology Web site maintained by Janet Hager: janet.hager@yale.edu

```
ATF     1
51      5
Type=GenePix ArrayList V1.0
BlockCount=48
BlockType=0
"Block1=500, 500, 100, 24, 175, 20, 175"
"Block2=4996, 500, 100, 24, 175, 20, 175"
"Block3=9492, 500, 100, 24, 175, 20, 175"
"Block4=13988, 500, 100, 24, 175, 20, 175"
"Block5=500, 4996, 100, 24, 175, 20, 175"
"Block6=4996, 4996, 100, 24, 175, 20, 175"
"Block7=9492, 4996, 100, 24, 175, 20, 175"
"Block8=13988, 4996, 100, 24, 175, 20, 175"
"Block9=500, 9492, 100, 24, 175, 20, 175"
"Block10=4996, 9492, 100, 24, 175, 20, 175"
"Block11=9492, 9492, 100, 24, 175, 20, 175"
"Block12=13988, 9492, 100, 24, 175, 20, 175"
"Block13=500, 13988, 100, 24, 175, 20, 175"
"Block14=4996, 13988, 100, 24, 175, 20, 175"
"Block15=9492, 13988, 100, 24, 175, 20, 175"
"Block16=13988, 13988, 100, 24, 175, 20, 175"
"Block17=500, 18484, 100, 24, 175, 20, 175"
"Block18=4996, 18484, 100, 24, 175, 20, 175"
"Block19=9492, 18484, 100, 24, 175, 20, 175"
"Block20=13988, 18484, 100, 24, 175, 20, 175"
"Block21=500, 22980, 100, 24, 175, 20, 175"
"Block22=4996, 22980, 100, 24, 175, 20, 175"
"Block23=9492, 22980, 100, 24, 175, 20, 175"
"Block24=13988, 22980, 100, 24, 175, 20, 175"
"Block25=500, 27476, 100, 24, 175, 20, 175"
"Block26=4996, 27476, 100, 24, 175, 20, 175"
"Block27=9492, 27476, 100, 24, 175, 20, 175"
"Block28=13988, 27476, 100, 24, 175, 20, 175"
"Block29=500, 31972, 100, 24, 175, 20, 175"
"Block30=4996, 31972, 100, 24, 175, 20, 175"
"Block31=9492, 31972, 100, 24, 175, 20, 175"
"Block32=13988, 31972, 100, 24, 175, 20, 175"
"Block33=500, 36468, 100, 24, 175, 20, 175"
"Block34=4996, 36468, 100, 24, 175, 20, 175"
"Block35=9492, 36468, 100, 24, 175, 20, 175"
"Block36=13988, 36468, 100, 24, 175, 20, 175"
"Block37=500, 40964, 100, 24, 175, 20, 175"
"Block38=4996, 40964, 100, 24, 175, 20, 175"
"Block39=9492, 40964, 100, 24, 175, 20, 175"
"Block40=13988, 40964, 100, 24, 175, 20, 175"
"Block41=500, 45460, 100, 24, 175, 20, 175"
"Block42=4996, 45460, 100, 24, 175, 20, 175"
"Block43=9492, 45460, 100, 24, 175, 20, 175"
"Block44=13988, 45460, 100, 24, 175, 20, 175"
"Block45=500, 49956, 100, 24, 175, 20, 175"
"Block46=4996, 49956, 100, 24, 175, 20, 175"
"Block47=9492, 49956, 100, 24, 175, 20, 175"
"Block48=13988, 49956, 100, 24, 175, 20, 175"
Block   Row     Column  ID      Name
1       1       1       NM_000911       "Opioid receptor, delta 1"
1       1       2       NM_001124       Adrenomedullin
1       1       3       NM_004024       Activating transcription factor 3
1       1       4       NM_002202       "ISL1 transcription factor, LIM/homeodomain,
1       1       5       NM_000384       Apolipoprotein B (including Ag(x) antigen)
1       1       6       NM_002306       "Lectin, galactoside-binding, soluble, 3 (ga
```

1	1	7	NM_000100	Cystatin B (stefin B)
1	1	8	NM_003250	"Thyroid hormone receptor, alpha (erythrobla
1	1	9	NM_000277	Phenylalanine hydroxylase
1	1	10	NM_000892	"Kallikrein B, plasma (Fletcher factor) 1"
1	1	11	NM_000180	"Guanylate cyclase 2D, membrane (retina-spec
1	1	12	NM_003140	Sex determining region Y
1	1	13	NM_000373	Uridine monophosphate synthetase (orotate ph
1	1	14	NM_000956	"Prostaglandin E receptor 2 (subtype EP2), 5
1	1	15	NM_000760	Colony stimulating factor 3 receptor (granul
1	1	16	NM_006685	Proline rich 3
1	1	17	NM_014403	"Sialyltransferase 7D ((alpha-N-acetylneuram
1	1	18	AL117595	Homo sapiens mRNA; cDNA DKFZp564C2063 (from
1	1	19	AK024956	"Homo sapiens cDNA: FLJ21303 fis, clone COL0
1	1	20	AL359592	Hypothetical protein DKFZp761H039
1	1	21	NM_001412	Eukaryotic translation initiation factor 1A
1	1	22	AL161960	Hypothetical protein FLJ21324
1	1	23	NM_006703	Nudix (nucleoside diphosphate linked moiety
1	1	24	AK055334	Hypothetical protein DKFZp566A1524
1	2	1	NM_003611	Oral-facial-digital syndrome 1
1	2	2	BC007933	Elongation factor for selenoprotein translat
1	2	3	AK000789	"Homo sapiens cDNA FLJ20782 fis, clone COL03
1	2	4	NM_004522	Kinesin family member 5C
1	2	5	NM_004870	Mannose-P-dolichol utilization defect 1
1	2	6	NM_030792	Hypothetical protein PP1665
1	2	7	NM_022735	Golgi phosphoprotein 1
1	2	8	NM_024533	Hypothetical protein FLJ22167
1	2	9	NM_020198	GK001 protein
1	2	10	AF035947	Cytokine inducible SH2-containing protein
1	2	11	AL137761	Homo sapiens mRNA; cDNA DKFZp586L2424 (from
1	2	12	AB046807	KIAA1587 protein
1	2	13	NM_007274	Cytosolic acyl coenzyme A thioester hydrolas
1	2	14	NM_024025	Hypothetical protein MGC1136
1	2	15	NM_003474	A disintegrin and metalloproteinase domain 1
1	2	16	NM_004871	Golgi SNAP receptor complex member 1
1	2	17	NM_017821	Hypothetical protein FLJ20435
1	2	18	NM_004768	"Splicing factor, arginine/serine-rich 11"
1	2	19	AB002295	KIAA0329 gene product
1	2	20	NM_004698	U4/U6-associated RNA splicing factor
1	2	21	NM_014400	GPI-anchored metastasis-associated protein h
1	2	22	NM_002940	"ATP-binding cassette, sub-family E (OABP),
1	2	23	NM_018127	ElaC homolog 2 (E. coli)
1	2	24	NM_021203	APMCF1 protein
1	3	1	NM_002163	Interferon consensus sequence binding protei
1	3	2	AK026873	"Homo sapiens, clone IMAGE:4431274, mRNA, pa
1	3	3	NM_001214	Chromosome 16 open reading frame 3
1	3	4	NM_014384	"Acyl-Coenzyme A dehydrogenase family, membe
1	3	5	AK054578	"Homo sapiens cDNA FLJ30016 fis, clone 3NB69
1	3	6	AB067801	Homolog of mouse quaking QKI (KH domain RNA
1	3	7	AK001740	Hypothetical protein DKFZp564O043
1	3	8	NM_007169	Phosphatidylethanolamine N-methyltransferase
1	3	9	AF231919	Chromosome 21 open reading frame 108
1	3	10	NM_015247	Cylindromatosis (turban tumor syndrome)
1	3	11	AB037725	KIAA1304 protein
1	3	12	NM_014824	KIAA0769 gene product
1	3	13	NM_002492	"NADH dehydrogenase (ubiquinone) 1 beta subc
1	3	14	BC009435	"Homo sapiens, Similar to RIKEN cDNA 2010317
1	3	15	NM_022335	Hypothetical protein PRO2849
1	3	16	NM_005001	"NADH dehydrogenase (ubiquinone) 1 alpha sub
1	3	17	AB033767	Chromosome 20 open reading frame 3
1	3	18	AB067485	KIAA1898 protein

1	3	19	AK056052	Homo sapiens clone CDABP0036 mRNA sequence
1	3	20	NM_018164	Hypothetical protein FLJ10637
1	3	21	NM_022762	Hypothetical protein FLJ22318
1	3	22	NM_014520	MYB binding protein (P160) 1a
1	3	23	NM_012261	Chromosome 20 open reading frame 103
1	3	24	AF339822	"Homo sapiens clone IMAGE:451939, mRNA seque
1	4	1	AL136835	Toll-interacting protein
1	4	2	NM_001170	Aquaporin 7
1	4	3	NM_001348	Death-associated protein kinase 3
1	4	4	NM_005851	Tumor suppressor deleted in oral cancer-rela
1	4	5	NM_005857	"Zinc metalloproteinase (STE24 homolog, yeas
1	4	6	NM_004638	HLA-B associated transcript 2
1	4	7	NM_007185	Trinucleotide repeat containing 4
1	4	8	BC006214	Hypothetical protein
1	4	9	NM_007216	Alpha integrin binding protein 63
1	4	10	NM_002401	Mitogen-activated protein kinase kinase kina
1	4	11	AK022156	"Homo sapiens cDNA FLJ12094 fis, clone HEMBB
1	4	12	BC015419	"Homo sapiens, clone MGC:21987 IMAGE:4396825
1	4	13	NM_024613	Phafin 2
1	4	14	NM_053043	Hypothetical protein MGC20460
1	4	15	NM_000112	"Solute carrier family 26 (sulfate transport
1	4	16	NM_020662	"MRS2-like, magnesium homeostasis factor (S.
1	4	17	NM_014795	Zinc finger homeobox 1b
1	4	18	NM_003368	Ubiquitin specific protease 1
1	4	19	AJ301564	Reserved
1	4	20	NM_002501	Nuclear factor I/X (CCAAT-binding transcript
1	4	21	NM_024320	Hypothetical protein MGC11242
1	4	22	NM_000204	I factor (complement)
1	4	23	NM_000519	"Hemoglobin, delta"
1	4	24	NM_001631	"Alkaline phosphatase, intestinal"
1	5	1	AK024472	Bcl-2 modifying factor
1	5	2	NM_018222	"Parvin, alpha"
1	5	3	NM_018965	Triggering receptor expressed on myeloid cel
1	5	4	NM_022449	"RAB17, member RAS oncogene family"
1	5	5	NM_002687	"Pinin, desmosome associated protein"
1	5	6	NM_017707	Hypothetical protein FLJ20199
1	5	7	NM_024637	"Beta-galactose-3-O-sulfotransferase, 4"
1	5	8	NM_024638	Hypothetical protein FLJ12960
1	5	9	NM_005894	CD5 antigen-like (scavenger receptor cystein
1	5	10	NM_032988	Transducin (beta)-like 2
1	5	11	AK026289	Likely ortholog of mouse lipoic acid synthas
1	5	12	NM_003381	Vasoactive intestinal peptide
1	5	13	NM_000579	Chemokine (C-C motif) receptor 5
1	5	14	NM_000352	"ATP-binding cassette, sub-family C (CFTR/MR
1	5	15	NM_006153	NCK adaptor protein 1
1	5	16	NM_021219	Junctional adhesion molecule 2
1	5	17	NM_022776	Oxysterol binding protein-like 11
1	5	18	AK057710	"Homo sapiens cDNA FLJ33148 fis, clone UTERU
1	5	19	AB051464	KIAA1677
1	5	20	NM_058229	"Homo sapiens cDNA FLJ32424 fis, clone SKMUS
1	5	21	NM_018992	Hypothetical protein
1	5	22	NM_006359	"Solute carrier family 9 (sodium/hydrogen ex
1	5	23	BC007910	"Homo sapiens cDNA FLJ25348 fis, clone TST01
1	5	24	AB037753	KIAA1332 protein
1	6	1	NM_000444	"Phosphate regulating gene with homologies t
1	6	2	NM_006891	"Crystallin, gamma D"
1	6	3	NM_002363	"Melanoma antigen, family B, 1"
1	6	4	NM_001351	Deleted in azoospermia-like
1	6	5	NM_024507	Hypothetical protein MGC10791
1	6	6	AK056582	"Homo sapiens cDNA FLJ32020 fis, clone NTONG

1	6	7	NM_002830	"Protein tyrosine phosphatase, non-receptor
1	6	8	NM_001685	"ATP synthase, H+ transporting, mitochondria
1	6	9	NM_004766	"Coatomer protein complex, subunit beta 2 (b
1	6	10	NM_000693	"Aldehyde dehydrogenase 1 family, member A3"
1	6	11	NM_004563	Phosphoenolpyruvate carboxykinase 2 (mitocho
1	6	12	NM_014761	KIAA0174 gene product
1	6	13	NM_002631	Phosphogluconate dehydrogenase
1	6	14	NM_015379	Brain protein I3
1	6	15	NM_004393	Dystroglycan 1 (dystrophin-associated glycop
1	6	16	NM_001920	Decorin
1	6	17	NM_001541	Heat shock 27kD protein 2
1	6	18	NM_002851	"Protein tyrosine phosphatase, receptor-type
1	6	19	NM_000251	"MutS homolog 2, colon cancer, nonpolyposis
1	6	20	NM_000709	"Branched chain keto acid dehydrogenase E1,
1	6	21	AK024911	Heterogeneous nuclear ribonucleoprotein M
1	6	22	NM_003168	Suppressor of Ty 4 homolog 1 (S. cerevisiae)
1	6	23	X72304	Corticotropin releasing hormone receptor 1
1	6	24	NM_005429	Vascular endothelial growth factor C
1	7	1	NM_001747	"Capping protein (actin filament), gelsolin-
1	7	2	AL133623	Similar to mouse Xrn1 / Dhm2 protein
1	7	3	NM_002822	Protein tyrosine kinase 9
1	7	4	AL117461	Homo sapiens mRNA; cDNA DKFZp586F1822 (from
1	7	5	AK025706	Adenosine monophosphate deaminase 2 (isoform
1	7	6	NM_004579	Mitogen-activated protein kinase kinase kina
1	7	7	NM_001760	Cyclin D3
1	7	8	NM_021077	Neuromedin B
1	7	9	NM_001213	Chromosome 1 open reading frame 1
1	7	10	NM_000523	Homeo box D13
1	7	11	NM_002910	Renin binding protein
1	7	12	NM_012452	Transmembrane activator and CAML interactor
1	7	13	AF052176	Homo sapiens clone 24457 mRNA sequence
1	7	14	NM_015904	Translation initiation factor IF2
1	7	15	NM_004309	Rho GDP dissociation inhibitor (GDI) alpha
1	7	16	NM_004202	"Thymosin, beta 4, Y chromosome"
1	7	17	NM_006231	"Polymerase (DNA directed), epsilon"
1	7	18	NM_005245	FAT tumor suppressor homolog 1 (Drosophila)
1	7	19	NM_015896	BLu protein
1	7	20	NM_021965	Phosphoglucomutase 5
1	7	21	NM_002302	Leukocyte cell-derived chemotaxin 2
1	7	22	NM_004818	"Prp28, U5 snRNP 100 kd protein"
1	7	23	AL080134	Homo sapiens mRNA; cDNA DKFZp434G043 (from c
1	7	24	U36759	"Human pre TCR alpha mRNA, partial cds"
1	8	1	NM_016607	ALEX3 protein
1	8	2	NM_013957	Neuregulin.1
1	8	3	AB012922	Metastasis-associated 1-like 1
1	8	4	NM_007106	Ubiquitin-like 3
1	8	5	NM_014633	KIAA0155 gene product
1	8	6	AB023148	KIAA0931 protein
1	8	7	AB032953	"Odd Oz/ten-m homolog 2 (Drosophila, mouse)"
1	8	8	NM_030939	Hypothetical protein FLJ12619
1	8	9	NM_004373	Cytochrome c oxidase subunit VIa polypeptide
1	8	10	NM_024835	C3HC4-type zinc finger protein
1	8	11	NM_001013	Ribosomal protein S9
1	8	12	AK023845	KIAA0729 protein
1	8	13	NM_005229	"ELK1, member of ETS oncogene family"
1	8	14	NM_058246	Similar to MRJ gene for a member of the DNAJ
1	8	15	AF188747	"Kallikrein 2, prostatic"
1	8	16	NM_018465	Uncharacterized hematopoietic stem/progenito
1	8	17	AK023151	Hypothetical protein FLJ13089
1	8	18	NM_000245	Met proto-oncogene (hepatocyte growth factor

1	8	19	AK027101	"Homo sapiens, clone IMAGE:3867243, mRNA"
1	8	20	NM_031946	"Centaurin, gamma 3"
1	8	21	NM_005761	Plexin C1
1	8	22	AK022434	"Homo sapiens cDNA FLJ12372 fis, clone MAMMA
1	8	23	AF264628	Uncharacterized gastric protein ZG24P
1	8	24	NM_021619	PR domain containing 12
1	9	1	NM_025072	Chromosome 9 open reading frame 15
1	9	2	NM_025075	Hypothetical protein FLJ23445
1	9	3	AB037726	KIAA1305 protein
1	9	4	BC011534	Hypothetical protein FLJ21128
1	9	5	AK021583	"Homo sapiens cDNA FLJ11521 fis, clone HEMBA
1	9	6	AB058725	KIAA1822 protein
1	9	7	AK024173	"Homo sapiens cDNA FLJ14111 fis, clone MAMMA
1	9	8	AK024152	"Homo sapiens cDNA FLJ14090 fis, clone MAMMA
1	9	9	AK056677	"Homo sapiens cDNA FLJ32115 fis, clone PANCR
1	9	10	AK021879	"Homo sapiens cDNA FLJ11817 fis, clone HEMBA
1	9	11	NM_022447	Topoisomerase-related function protein 4-2
1	9	12	AB058708	KIAA1805 protein
1	9	13	AK055808	"Homo sapiens, clone MGC:27123 IMAGE:4793294
1	9	14	NM_025165	Hypothetical protein FLJ22637
1	9	15	NM_003804	Receptor (TNFRSF)-interacting serine-threoni
1	9	16	AJ007557	"Potassium inwardly-rectifying channel, subf
1	9	17	AK024109	"Homo sapiens cDNA FLJ14047 fis, clone HEMBA
1	9	18	AL110180	Homo sapiens mRNA; cDNA DKFZp566C034 (from c
1	9	19	NM_002235	"Potassium voltage-gated channel, shaker-rel
1	9	20	AB046860	Hypothetical protein FLJ20753
1	9	21	AK021796	"Homo sapiens cDNA FLJ11734 fis, clone HEMBA
1	9	22	AK022401	"Homo sapiens cDNA FLJ12339 fis, clone MAMMA
1	9	23	AK021538	"Homo sapiens cDNA FLJ11476 fis, clone HEMBA
1	9	24	AB028947	KIAA1024 protein
1	10	1	NM_002196	Insulinoma-associated 1
1	10	2	NM_002841	"Protein tyrosine phosphatase, receptor type
1	10	3	NM_001869	Carboxypeptidase A2 (pancreatic)
1	10	4	NM_000812	"Gamma-aminobutyric acid (GABA) A receptor,
1	10	5	BC012356	Adenylosuccinate synthase
1	10	6	NM_003244	TGFB-induced factor (TALE family homeobox)
1	10	7	AB020689	KIAA0882 protein
1	10	8	NM_006415	"Serine palmitoyltransferase, long chain bas
1	10	9	AB023230	KIAA1013 protein
1	10	10	NM_019034	"Ras homolog gene family, member F (in filop
1	10	11	NM_024343	Hypothetical protein MGC10764
1	10	12	NM_000734	"CD3Z antigen, zeta polypeptide (TiT3 comple
1	10	13	NM_019033	Hypothetical protein
1	10	14	AB033083	KIAA1257 protein
1	10	15	NM_033123	Testis-development related NYD-SP27
1	10	16	NM_003025	SH3-domain GRB2-like 1
1	10	17	AL117636	Homo sapiens mRNA; cDNA DKFZp434H205 (from c
1	10	18	NM_015112	KIAA0807 protein
1	10	19	NM_004785	"Solute carrier family 9 (sodium/hydrogen ex
1	10	20	AB033019	KIAA1193 protein
1	10	21	NM_017701	Rho GTPase activating protein 8
1	10	22	NM_030935	TSC-22-like
1	10	23	AB067479	DKFZP434O125 protein
1	10	24	AK027705	KIAA1275 protein
1	11	1	NM_015921	Divalent cation tolerant protein CUTA
1	11	2	AB037799	KIAA1378 protein
1	11	3	BC013592	"Homo sapiens, Similar to RIKEN cDNA 1700037
1	11	4	NM_022831	Hypothetical protein FLJ12806
1	11	5	NM_002017	Friend leukemia virus integration 1
1	11	6	NM_024718	Hypothetical protein FLJ10101

1	11	7	NM_016022	CGI-78 protein
1	11	8	AL136640	PABP-interacting protein 2
1	11	9	AF119818	"Discs, large (Drosophila) homolog-associate
1	11	10	NM_016592	GNAS complex locus
1	11	11	NM_014241	"Protein tyrosine phosphatase-like (proline
1	11	12	NM_003506	Frizzled homolog 6 (Drosophila)
1	11	13	BE884686	"ESTs, Highly similar to LB4D_HUMAN NADP-DEP
1	11	14	NM_033266	ER to nucleus signalling 2
1	11	15	NM_003929	"RAB7, member RAS oncogene family-like 1"
1	11	16	NM_024052	Hypothetical protein MGC3048
1	11	17	NM_014487	Nucleolar cysteine-rich protein
1	11	18	AK024007	"Homo sapiens cDNA FLJ13945 fis, clone Y79AA
1	11	19	AK057278	Contactin 4
1	11	20	AK025222	"Homo sapiens cDNA: FLJ21569 fis, clone COL0
1	11	21	U50277	"Human breast cancer suppressor element Ishmael Uppe
1	11	22	NM_000219	"Potassium voltage-gated channel, Isk-relate
1	11	23	NM_004967	"Integrin-binding sialoprotein (bone sialopr
1	11	24	AK022293	"Homo sapiens cDNA FLJ12231 fis, clone MAMMA
1	12	1	AL109699	Homo sapiens mRNA full length insert cDNA cl
1	12	2	AK056381	"Homo sapiens cDNA FLJ31819 fis, clone NT2RP
1	12	3	BC015422	"Homo sapiens, clone MGC:21990 IMAGE:4397794
1	12	4	BC014917	"Homo sapiens, Similar to kinase interacting
1	12	5	NM_016239	Myosin XVA
1	12	6	NM_018515	Hypothetical protein PRO2176
1	12	7	AF070674	Baculoviral IAP repeat-containing 3
1	12	8	AL133102	CGI-142
1	12	9	NM_006194	Paired box gene 9
1	12	10	BG163465	ESTs
1	12	11	NM_014666	KIAA0171 gene product
1	12	12	AB051461	KIAA1674
1	12	13	AK025924	"Homo sapiens cDNA: FLJ22271 fis, clone HRC0
1	12	14	AL050035	Homo sapiens mRNA; cDNA DKFZp566H0124 (from
1	12	15	AL080095	Homo sapiens mRNA; cDNA DKFZp564O0862 (from
1	12	16	NM_014489	FGF receptor activating protein 1
1	12	17	NM_003181	"T, brachyury homolog (mouse)"
1	12	18	NM_033224	Purine-rich element binding protein B
1	12	19	NM_024509	Hypothetical protein MGC2656
1	12	20	AF301223	Homo sapiens chromosome 18 unknown mRNA sequ
1	12	21	NM_018166	Hypothetical protein FLJ10647
1	12	22	NM_015653	DKFZP566F0546 protein
1	12	23	AF384667	"Homo sapiens carboxypeptidase A5 mRNA, comp
1	12	24	AL096732	Homo sapiens mRNA; cDNA DKFZp434N074 (from c
1	13	1	AF052119	Homo sapiens clone 23622 mRNA sequence
1	13	2	NM_014681	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide
1	13	3	NM_000775	"Cytochrome P450, subfamily IIJ (arachidonic
1	13	4	NM_003468	Frizzled homolog 5 (Drosophila)
1	13	5	NM_024519	Hypothetical protein FLJ13725
1	13	6	AP001753	"Homo sapiens genomic DNA, chromosome 21q, s
1	13	7	AF311312	Sperm associated antigen 1
1	13	8	NM_001444	Fatty acid binding protein 5 (psoriasis-asso
1	13	9	AK024299	"Homo sapiens cDNA FLJ14237 fis, clone NT2RP
1	13	10	NM_005994	T-box 2
1	13	11	AA278251	ESTs
1	13	12	AW374095	ESTs
1	13	13	NM_018399	Vanin 3
1	13	14	M26880	Ubiquitin C
1	13	15	NM_007352	"Elastase 3B, pancreatic"
1	13	16	NM_018961	"Ubiquitin associated and SH3 domain contain
1	13	17	AC007059	Hypothetical protein DKFZp547M136 similar to
1	13	18	AB028992	KIAA1069 protein

1	13	19	AF349446	"Solute carrier family 14 (urea transporter)
1	13	20	AL133663	Homo sapiens mRNA; cDNA DKFZp434O1521 (from
1	13	21	NM_017935	Hypothetical protein FLJ20706
1	13	22	NM_014668	KIAA0575 gene product
1	13	23	NM_031464	"Hypothetical protein MGC11287 similar to ri
1	13	24	NM_004101	Coagulation factor II (thrombin) receptor-li
1	14	1	AK056862	M-phase phosphoprotein 10 (U3 small nucleola
1	14	2	NM_004098	Empty spiracles homolog 2 (Drosophila)
1	14	3	NM_006552	Lipophilin A (uteroglobin family member)
1	14	4	AB029025	KIAA1102 protein
1	14	5	NM_002069	"Guanine nucleotide binding protein (G prote
1	14	6	NM_006669	"Leukocyte immunoglobulin-like receptor, sub
1	14	7	AK023614	"Homo sapiens cDNA FLJ13552 fis, clone PLACE
1	14	8	AB033037	KIAA1211 protein
1	14	9	NM_004256	Organic cationic transporter-like 3
1	14	10	NM_006707	Butyrophilin-like 3
1	14	11	AL050185	Homo sapiens mRNA; cDNA DKFZp586A0423 (from
1	14	12	AL080137	DKFZP434J193 protein
1	14	13	AK025163	"Homo sapiens cDNA: FLJ21510 fis, clone COL0
1	14	14	AF131837	Homo sapiens clone 24881 mRNA sequence
1	14	15	NM_014977	KIAA0670 protein/acinus
1	14	16	AK024217	U6 snRNA-associated Sm-like protein
1	14	17	NM_004949	Desmocollin 2
1	14	18	AF308287	Serologically defined breast cancer antigen
1	14	19	AK027191	"Homo sapiens cDNA: FLJ23538 fis, clone LNG0
1	14	20	AF415175	DKFZP434D146 protein
1	14	21	NM_012285	"Potassium voltage-gated channel, subfamily
1	14	22	AL109783	Homo sapiens mRNA full length insert cDNA cl
1	14	23	AF444143	Spastic paraplegia 3A (autosomal dominant)
1	14	24	NM_004133	"Hepatocyte nuclear factor 4, gamma"
1	15	1	NM_005260	Growth differentiation factor 9
1	15	2	NM_004248	G protein-coupled receptor 10
1	15	3	NM_005549	"Potassium voltage-gated channel, shaker-rel
1	15	4	NM_005958	Melatonin receptor 1A
1	15	5	NM_003555	"Olfactory receptor, family 1, subfamily G,
1	15	6	NM_003771	"Keratin, hair, acidic, 6"
1	15	7	NM_012114	"Caspase 14, apoptosis-related cysteine prot
1	15	8	AC004983	Homo sapiens PAC clone RP5-1163J12 from 7q21
1	15	9	AK055458	"Homo sapiens cDNA FLJ30757 fis, clone FEBRA
1	15	10	AK001120	"Homo sapiens cDNA FLJ10258 fis, clone HEMBB
1	15	11	NM_012152	"Endothelial differentiation, lysophosphatid
1	15	12	AB037790	Heme-regulated initiation factor 2-alpha kin
1	15	13	L05500	Adenylate cyclase 1 (brain)
1	15	14	NM_004321	Axonal transport of synaptic vesicles
1	15	15	AK026460	"Homo sapiens cDNA: FLJ22807 fis, clone KAIA
1	15	16	NM_021031	Cytochrome c-like antigen
1	15	17	NM_018382	Hypothetical protein FLJ11292
1	15	18	AB037738	KIAA1317 protein
1	15	19	AI476463	EST
1	15	20	AL122100	Homo sapiens mRNA; cDNA DKFZp434D0617 (from
1	15	21	AL109809	Human DNA sequence from clone RP4-673D20 on
1	15	22	NM_017416	Interleukin 1 receptor accessory protein-lik
1	15	23	NM_014627	G protein-coupled receptor 57
1	15	24	NM_023919	"Taste receptor, type 2, member 7"
1	16	1	AK027646	Hypothetical protein
1	16	2	NM_032644	Hypothetical protein MGC2452
1	16	3	NM_032237	Hypothetical protein FLJ23356
1	16	4	NM_006389	Oxygen regulated protein (150kD)
1	16	5	AK057820	"Unactive progesterone receptor, 23 kD"
1	16	6	BC018205	"Homo sapiens, clone IMAGE:3464710, mRNA, pa

1	16	7	AF304052	DKFZP586G1122 protein
1	16	8	NM_005138	SCO cytochrome oxidase deficient homolog 2 (
1	16	9	AK026416	"Homo sapiens cDNA: FLJ22763 fis, clone KAIA
1	16	10	NM_019084	Hypothetical protein FLJ10895
1	16	11	AK024984	"Homo sapiens cDNA: FLJ21331 fis, clone COL0
1	16	12	AK025490	"Homo sapiens cDNA: FLJ21837 fis, clone HEP0
1	16	13	NM_000157	"Glucosidase, beta; acid (includes glucosylc
1	16	14	NM_018958	Chromosome 15 open reading frame 2
1	16	15	NM_018517	Hypothetical protein PRO2214
1	16	16	NM_018532	Hypothetical protein PRO2610
1	16	17	AK022106	"Homo sapiens cDNA FLJ12044 fis, clone HEMBB
1	16	18	NM_031934	"RAB34, member RAS oncogene family"
1	16	19	AB055890	Lymphoid blast crisis oncogene
1	16	20	NM_000561	Glutathione S-transferase M1
1	16	21	NM_021828	Heparanase-like protein
1	16	22	BC012899	"Homo sapiens, clone MGC:18222 IMAGE:4156395
1	16	23	AL049244	Homo sapiens mRNA; cDNA DKFZp564C163 (from c
1	16	24	AL049283	Homo sapiens mRNA; cDNA DKFZp564M163 (from c
1	17	1	AL137403	Homo sapiens mRNA; cDNA DKFZp434L092 (from c
1	17	2	AL137486	Homo sapiens mRNA; cDNA DKFZp434L1230 (from
1	17	3	AL157472	Homo sapiens mRNA; cDNA DKFZp761J2024 (from
1	17	4	AL162077	Homo sapiens mRNA; cDNA DKFZp761A219 (from c
1	17	5	AF172327	"Homo sapiens clone 709724 unknown mRNA, com
1	17	6	AF090945	Homo sapiens clone HQ0670
1	17	7	AK021523	"Homo sapiens cDNA FLJ11461 fis, clone HEMBA
1	17	8	AK021652	"Homo sapiens cDNA FLJ11590 fis, clone HEMBA
1	17	9	NM_031854	Keratin associated protein 4.12
1	17	10	NM_033185	Keratin associated protein 3.3
1	17	11	AL512755	Homo sapiens mRNA; cDNA DKFZp667K226 (from c
1	17	12	AL512720	Hypothetical protein DKFZp547J222
1	17	13	AF063599	"Homo sapiens brain my041 protein mRNA, comp
1	17	14	AL050335	"Human DNA sequence from clone RP1-190J20 on
1	17	15	Z22780	"Cylicin, basic protein of sperm head cytoskeleton 1
1	17	16	AF288406	Homo sapiens G protein interaction factor 2-
1	17	17	AL133023	Hypothetical protein DKFZp434A1022
1	17	18	NM_031210	Hypothetical protein DC50
1	17	19	NM_024016	Homeo box B8
1	17	20	NM_018172	Hypothetical protein FLJ10661
1	17	21	AF143869	Homo sapiens clone IMAGE:111469 mRNA sequenc
1	17	22	NM_005699	Interleukin 18 binding protein
1	17	23	AK057139	Hypothetical protein MGC3731
1	17	24	AK055779	V-akt murine thymoma viral oncogene homolog
1	18	1	NM_004757	"Small inducible cytokine subfamily E, membe
1	18	2	AF177377	Chromosome 2 open reading frame 2
1	18	3	BC006119	"Homo sapiens, clone IMAGE:3505629, mRNA, pa
1	18	4	NM_032346	Hypothetical protein MGC13096
1	18	5	AJ301580	DMRT-like family A2
1	18	6	NM_032363	HEIL2 protein
1	18	7	NM_032805	Hypothetical protein FLJ14549
1	18	8	NM_032890	Hypothetical protein MGC13130
1	18	9	AK023586	"Homo sapiens cDNA FLJ13524 fis, clone PLACE
1	18	10	NM_018001	Hypothetical protein FLJ10120
1	18	11	BC009775	"Homo sapiens, Similar to hypothetical prote
1	18	12	NM_006587	Corin
1	18	13	NM_006051	FE65-LIKE 2
1	18	14	NM_003917	"Adaptor-related protein complex 1, gamma 2
1	18	15	NM_005918	"Malate dehydrogenase 2, NAD (mitochondrial)
1	18	16	D87438	KIAA0251 protein
1	18	17	NM_000850	Glutathione S-transferase M4
1	18	18	NM_002643	"Phosphatidylinositol glycan, class F"

1	18	19	AB067474	KIAA1887 protein
1	18	20	AK023083	"Homo sapiens cDNA FLJ13021 fis, clone NT2RP
1	18	21	BC014346	"Homo sapiens, clone IMAGE:4042988, mRNA, pa
1	18	22	AF411191	"Homo sapiens SERPINB12 (SERPINB12) mRNA, co
1	18	23	BC015684	"Homo sapiens, Similar to Sjogren syndrome a
1	18	24	BC015416	"Homo sapiens, Similar to hypothetical prote
1	19	1	AK057915	"Homo sapiens cDNA FLJ25186 fis, clone CBR09
1	19	2	AK057861	"Homo sapiens cDNA FLJ25132 fis, clone CBR06
1	19	3	AK057679	"Homo sapiens cDNA FLJ33117 fis, clone TRACH
1	19	4	AK057638	"Homo sapiens cDNA FLJ33076 fis, clone TRACH
1	19	5	AK057513	"Homo sapiens cDNA FLJ32951 fis, clone TESTI
1	19	6	AK057458	"Homo sapiens cDNA FLJ32896 fis, clone TESTI
1	19	7	AK057355	"Homo sapiens cDNA FLJ32793 fis, clone TESTI
1	19	8	AK057320	"Homo sapiens cDNA FLJ32758 fis, clone TESTI
1	19	9	BC011816	"Homo sapiens, clone IMAGE:3346747, mRNA, pa
1	19	10	AK054757	"Homo sapiens cDNA FLJ30195 fis, clone BRACE
1	19	11	M54968	V-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homol
1	19	12	NM_022782	M-phase phosphoprotein 9
1	19	13	AK056070	"Homo sapiens cDNA FLJ31508 fis, clone NT2NE
1	19	14	AK057161	"Homo sapiens cDNA FLJ32599 fis, clone STOMA
1	19	15	AK056250	"Homo sapiens cDNA FLJ31688 fis, clone NT2RI
1	19	16	NM_012227	Pseudoautosomal GTP-binding protein-like
1	19	17	BLANK	BLANK
1	19	18	BLANK	BLANK
1	19	19	BLANK	BLANK
1	19	20	BLANK	BLANK
1	19	21	BLANK	BLANK
1	19	22	BLANK	BLANK
1	19	23	BLANK	BLANK
1	19	24	BLANK	BLANK
1	20	1	NM_001967	"Eukaryotic translation initiation factor 4A
1	20	2	NM_006082	"Tubulin, alpha, ubiquitous"
1	20	3	NM_001967	"Eukaryotic translation initiation factor 4A
1	20	4	NM_006082	"Tubulin, alpha, ubiquitous"
1	20	5	NM_001967	"Eukaryotic translation initiation factor 4A
1	20	6	NM_006082	"Tubulin, alpha, ubiquitous"
1	20	7	NM_001967	"Eukaryotic translation initiation factor 4A
1	20	8	NM_006082	"Tubulin, alpha, ubiquitous"
2	1	1	NM_000823	Growth hormone releasing hormone receptor
2	1	2	NM_000419	"Integrin, alpha 2b (platelet glycoprotein I
2	1	3	NM_000191	3-hydroxymethyl-3-methylglutaryl-Coenzyme A
2	1	4	NM_001557	"Interleukin 8 receptor, beta"
2	1	5	NM_003143	Single-stranded DNA binding protein
2	1	6	NM_017534	"Myosin, heavy polypeptide 2, skeletal muscl
2	1	7	NM_000316	Parathyroid hormone receptor 1
2	1	8	NM_003141	"Sjogren syndrome antigen A1 (52kD, ribonucl
2	1	9	NM_000638	"Vitronectin (serum spreading factor, somato
2	1	10	NM_002950	Ribophorin I
2	1	11	NM_001187	B melanoma antigen
2	1	12	NM_003241	Transglutaminase 4 (prostate)
2	1	13	NM_014878	KIAA0020 gene product
2	1	14	NM_005137	DiGeorge syndrome critical region gene 2
2	1	15	NM_000283	"Phosphodiesterase 6B, cGMP-specific, rod, b
2	1	16	NM_001958	Eukaryotic translation elongation factor 1 a
2	1	17	D79998	KIAA0176 protein
2	1	18	NM_004518	"Potassium voltage-gated channel, KQT-like s
2	1	19	AK021776	"Homo sapiens cDNA FLJ11714 fis, clone HEMBA
2	1	20	NM_015997	CGI-41 protein
2	1	21	NM_006391	RAN binding protein 7
2	1	22	NM_016397	TH1-like (Drosophila)

2	1	23	NM_006698	Bladder cancer associated protein
2	1	24	BC014900	Hypothetical protein
2	2	1	AJ238403	Huntingtin interacting protein B
2	2	2	AK025697	"Homo sapiens cDNA: FLJ22044 fis, clone HEP0
2	2	3	NM_015701	Hypothetical protein
2	2	4	NM_007281	Scrapie responsive protein 1
2	2	5	AK024428	"Pleckstrin homology, Sec7 and coiled/coil d
2	2	6	AK024396	Acetyl-Coenzyme A synthetase 2 (AMP forming)
2	2	7	NM_032442	G protein pathway suppressor 2
2	2	8	AF101051	Claudin 1
2	2	9	NM_032208	Tumor endothelial marker 8
2	2	10	NM_001897	Chondroitin sulfate proteoglycan 4 (melanoma
2	2	11	NM_032305	Hypothetical protein MGC3200
2	2	12	NM_016172	Putative glialblastoma cell differentiation-
2	2	13	AK022487	"Homo sapiens cDNA FLJ12425 fis, clone MAMMA
2	2	14	NM_006480	Regulator of G-protein signalling 14
2	2	15	NM_016085	Apoptosis related protein APR-3
2	2	16	NM_005096	Zinc finger protein 261
2	2	17	AB032989	KIAA1163 protein
2	2	18	NM_003170	Suppressor of Ty 6 homolog (S. cerevisiae)
2	2	19	NM_015270	Adenylate cyclase 6
2	2	20	AK057657	"Homo sapiens cDNA FLJ33095 fis, clone TRACH
2	2	21	NM_018686	CMP-N-acetylneuraminic acid synthase
2	2	22	AF007149	Homo sapiens clone 24771 mRNA sequence
2	2	23	NM_017691	Hypothetical protein FLJ20156
2	2	24	AK023168	KIAA0678 protein
2	3	1	BC008810	"Homo sapiens, clone IMAGE:3948909, mRNA, pa
2	3	2	BC007548	"Homo sapiens, clone IMAGE:2959994, mRNA"
2	3	3	BC013747	"Homo sapiens, clone IMAGE:3869112, mRNA"
2	3	4	AB023166	"Citron (rho-interacting, serine/threonine k
2	3	5	AK002211	KIAA1272 protein
2	3	6	NM_032324	Hypothetical protein MGC13186
2	3	7	NM_002873	RAD17 homolog (S. pombe)
2	3	8	NM_017904	Hypothetical protein FLJ20619
2	3	9	NM_001902	Cystathionase (cystathionine gamma-lyase)
2	3	10	AL137679	Homo sapiens mRNA; cDNA DKFZp434D2426 (from
2	3	11	NM_017593	Homolog of mouse BMP-2 inducible kinase
2	3	12	NM_005672	Prostate stem cell antigen
2	3	13	AB032968	P21(CDKN1A)-activated kinase 4
2	3	14	AK025464	"Homo sapiens cDNA: FLJ21811 fis, clone HEP0
2	3	15	NM_016593	Oxysterol 7alpha-hydroxylase
2	3	16	NM_005882	Macrophage erythroblast attacher
2	3	17	NM_013366	Anaphase-promoting complex subunit 2
2	3	18	NM_003492	Chromosome X open reading frame 12
2	3	19	NM_005867	Down syndrome critical region gene 4
2	3	20	AB002365	KIAA0367 protein
2	3	21	AK024433	Mitochondrial ribosomal protein S25
2	3	22	AK055857	"Homo sapiens cDNA FLJ31295 fis, clone KIDNE
2	3	23	NM_016542	Serine/threonine protein kinase MASK
2	3	24	AF070632	Homo sapiens clone 24405 mRNA sequence
2	4	1	AK056372	"Homo sapiens cDNA FLJ31810 fis, clone NT2RI
2	4	2	AL050166	Homo sapiens mRNA; cDNA DKFZp586D1122 (from
2	4	3	NM_001382	Dolichyl-phosphate (UDP-N-acetylglucosamine)
2	4	4	NM_005503	"Amyloid beta (A4) precursor protein-binding
2	4	5	AF426160	Histone deacetylase 10
2	4	6	NM_052880	Hypothetical protein MGC17330
2	4	7	NM_014409	"TAF5-like RNA polymerase II, p300/CBP-assoc
2	4	8	NM_020147	Hypothetical protein from EUROIMAGE 511235
2	4	9	NM_018252	Hypothetical protein FLJ10874
2	4	10	AL136599	Sentrin/SUMO-specific protease

2	4	11	U10991 G2 protein
2	4	12	AF302494 "Potassium voltage-gated channel, Isk-relate
2	4	13	NM_004718 Cytochrome c oxidase subunit VIIa polypeptid
2	4	14	NM_018022 Hypothetical protein FLJ10199
2	4	15	NM_003098 "Syntrophin, alpha 1 (dystrophin-associated
2	4	16	AK026883 "Homo sapiens cDNA: FLJ23230 fis, clone CAE0
2	4	17	NM_006006 "Zinc finger protein 145 (Kruppel-like, expr
2	4	18	NM_006562 Transcription factor similar to D. melanogas
2	4	19	BC007069 Hypothetical protein FLJ11149
2	4	20	NM_002889 Retinoic acid receptor responder (tazarotene
2	4	21	NM_002285 Lymphoid nuclear protein related to AF4
2	4	22	BC010563 "Homo sapiens, clone MGC:18111 IMAGE:4152811
2	4	23	NM_004914 "RAB36, member RAS oncogene family"
2	4	24	AK057716 Hypothetical protein MGC11115
2	5	1	NM_000334 "Sodium channel, voltage-gated, type IV, alp
2	5	2	S73288 "Small proline-rich protein SPRK [human, odontogenic
2	5	3	NM_032131 Hypothetical protein DKFZp434P0714
2	5	4	BC012267 "Homo sapiens, clone IMAGE:3866403, mRNA"
2	5	5	AL157461 Homo sapiens mRNA; cDNA DKFZp434K152 (from c
2	5	6	AK057235 Hypothetical protein DKFZp434B217
2	5	7	AK056606 "Homo sapiens cDNA FLJ32044 fis, clone NTONG
2	5	8	AK024874 "Homo sapiens cDNA: FLJ21221 fis, clone COL0
2	5	9	NM_018468 Uncharacterized hematopoietic stem/progenito
2	5	10	NM_022772 Hypothetical protein FLJ21935
2	5	11	NM_016489 Uridine 5' monophosphate hydrolase 1
2	5	12	NM_033448 Keratin 6 irs
2	5	13	NM_017897 Hypothetical protein FLJ20604
2	5	14	NM_006856 Activating transcription factor 7
2	5	15	NM_003149 Src homology three (SH3) and cysteine rich d
2	5	16	AB011539 "EGF-like-domain, multiple 3"
2	5	17	NM_021925 Hypothetical protein FLJ21820
2	5	18	NM_002831 "Protein tyrosine phosphatase, non-receptor
2	5	19	NM_014772 KIAA0427 gene product
2	5	20	NM_003183 "A disintegrin and metalloproteinase domain
2	5	21	AL133662 KIAA0913 protein
2	5	22	NM_006033 "Lipase, endothelial"
2	5	23	AK055310 "Homo sapiens cDNA FLJ30748 fis, clone FEBRA
2	5	24	BC009231 "Homo sapiens, clone IMAGE:3959816, mRNA, pa
2	6	1	NM_000904 "NAD(P)H dehydrogenase, quinone 2"
2	6	2	NM_003460 Zona pellucida glycoprotein 2 (sperm recepto
2	6	3	NM_016732 "RNA binding protein (autoantigenic, hnRNP-a
2	6	4	NM_006827 Transmembrane trafficking protein
2	6	5	NM_012381 "Origin recognition complex, subunit 3-like
2	6	6	NM_001906 Chymotrypsinogen B1
2	6	7	D87452 KIAA0263 gene product
2	6	8	NM_003257 Tight junction protein 1 (zona occludens 1)
2	6	9	NM_001748 "Calpain 2, (m/II) large subunit"
2	6	10	NM_005380 "Neuroblastoma, suppression of tumorigenicit
2	6	11	NM_000506 Coagulation factor II (thrombin)
2	6	12	NM_006099 Protein inhibitor of activated STAT3
2	6	13	NM_006741 "Protein phosphatase 1, regulatory (inhibito
2	6	14	NM_000354 "Serine (or cysteine) proteinase inhibitor,
2	6	15	NM_004359 Cell division cycle 34
2	6	16	NM_001006 Ribosomal protein S3A
2	6	17	D86985 KIAA0232 gene product
2	6	18	NM_001968 Eukaryotic translation initiation factor 4E
2	6	19	X76061 Retinoblastoma-like 2 (p130)
2	6	20	NM_012198 "Grancalcin, EF-hand calcium binding protein
2	6	21	AB011154 KIAA0582 protein
2	6	22	NM_001783 CD79A antigen (immunoglobulin-associated alp

2	6	23	NM_012329	Monocyte to macrophage differentiation-assoc
2	6	24	U79242	Human clone 23560 mRNA sequence
2	7	1	NM_000227	"Laminin, alpha 3 (nicein (150kD), kalinin (
2	7	2	X59405	"Membrane cofactor protein (CD46, trophoblast-lympho
2	7	3	NM_003282	"Troponin I, skeletal, fast"
2	7	4	NM_001914	Cytochrome b-5
2	7	5	NM_024668	Hypothetical protein FLJ20288
2	7	6	NM_032343	Hypothetical protein MGC13016
2	7	7	NM_015343	Hypothetical protein
2	7	8	NM_002494	"NADH dehydrogenase (ubiquinone) 1, subcompl
2	7	9	NM_004680	"Chromodomain protein, Y chromosome, 1"
2	7	10	AK025225	"Homo sapiens cDNA: FLJ21572 fis, clone COL0
2	7	11	NM_016950	Testican 3
2	7	12	AC007956	ZAP3 protein
2	7	13	NM_001585	Chromosome 22 open reading frame 1
2	7	14	NM_016952	Cell adhesion molecule-related/down-regulate
2	7	15	NM_019598	"Homo sapiens kallikrein 12 (KLK12), mRNA"
2	7	16	NM_031302	Gycosyltransferase
2	7	17	NM_003836	Delta-like 1 homolog (Drosophila)
2	7	18	BC011864	"Homo sapiens, Similar to lung carcinoma myc
2	7	19	NM_014642	KIAA0036 gene product
2	7	20	NM_004742	BAI1-associated protein 1
2	7	21	NM_023080	Hypothetical protein FLJ20989
2	7	22	AK055529	"Homo sapiens cDNA FLJ30967 fis, clone HEART
2	7	23	NM_003318	TTK protein kinase
2	7	24	NM_019105	Tenascin XB
2	8	1	AB037721	KIAA1300 protein
2	8	2	NM_001967	"Eukaryotic translation initiation factor 4A
2	8	3	NM_014501	Ubiquitin carrier protein
2	8	4	NM_005211	"Colony stimulating factor 1 receptor, forme
2	8	5	D26067	KIAA0033 protein
2	8	6	AK024300	"Homo sapiens cDNA FLJ14238 fis, clone NT2RP
2	8	7	NM_003624	RAN binding protein 3
2	8	8	AB067512	KIAA1925 protein
2	8	9	AK023911	"Homo sapiens cDNA FLJ13849 fis, clone THYRO
2	8	10	NM_022098	Hypothetical protein LOC63929
2	8	11	NM_012153	Ets homologous factor
2	8	12	NM_004435	Endonuclease G
2	8	13	NM_021082	"Solute carrier family 15 (H+/peptide transp
2	8	14	NM_000578	"Solute carrier family 11 (proton-coupled di
2	8	15	AK025582	Guanine nucleotide binding protein (G protei
2	8	16	AF204231	Golgin-67
2	8	17	AK021524	"Homo sapiens cDNA FLJ11462 fis, clone HEMBA
2	8	18	NM_032446	MEGF10 protein
2	8	19	AK021795	"Homo sapiens cDNA FLJ11733 fis, clone HEMBA
2	8	20	NM_024973	Hypothetical protein FLJ11781
2	8	21	NM_024977	Hypothetical protein FLJ12078
2	8	22	AK022213	Hypothetical protein FLJ12151
2	8	23	AK022395	"Homo sapiens cDNA FLJ12333 fis, clone MAMMA
2	8	24	AK022435	"Homo sapiens cDNA FLJ12373 fis, clone MAMMA
2	9	1	NM_025105	Hypothetical protein FLJ12409
2	9	2	NM_003420	Zinc finger protein 35 (clone HF.10)
2	9	3	NM_025117	Hypothetical protein FLJ11871
2	9	4	AK025909	"Homo sapiens cDNA: FLJ22256 fis, clone HRC0
2	9	5	NM_025129	Hypothetical protein FLJ22688
2	9	6	BC012337	Hypothetical protein FLJ22761
2	9	7	AK022239	"Homo sapiens cDNA FLJ12177 fis, clone MAMMA
2	9	8	AK024308	"Homo sapiens cDNA FLJ14246 fis, clone OVARC
2	9	9	AK000864	"Homo sapiens cDNA FLJ10002 fis, clone HEMBA
2	9	10	AK000966	"Homo sapiens cDNA FLJ10104 fis, clone HEMBA

Welcome to VIB MicroArray Facility !

In April 1999, the Flanders Interuniversity Institute for Biotechnology (VIB) initiated a core facility for cDNA MicroArray analysis. The facility became operational in December 1999 and is located at Gasthuisberg, Leuven–Belgium. The purpose of the facility is to provide researchers access to a full MicroArray analysis service that includes:

- Dedicated Service Facility for MicroArray Analysis
- Advice in Experimental set-up
- Advice in choosing the appropriate technology platform
- Quality check of RNA integrity and purity
- Probe labeling starting from 5µg down to 100ng total RNA
- Quality check of labeled material
- Hybridisation, Scanning and Image Analysis
- Data pre-processing
- MicroArray follow-up with customized small size cDNA-arrays
- Advice in Data Analysis

The Facility supports 4 different MicroArray platforms:

- VIB In-House cDNA and Oligo-MicroArrays
- Affymetrix, GeneChip®
- Agilent Technologies, Agilent SurePrint
- Amersham Biosciences, CodeLink™

(CIU=HS4)	SPOTLABEL		MAF_CLONE_ID	SOURCE_CLONE_ID	CLONE_ACCESSION	SHORT_NAME
1	250001	cYIR01	Calib 01	"ratio 1:1"	APB	Calib 01
2	250002	cYIR02	Calib 02	"ratio 1:1"	APB	Calib 02
3	250003	cYIR03	Calib 03	"ratio 1:1"	APB	Calib 03
4	250004	cYIR04	Calib 04	"ratio 1:1"	APB	Calib 04
5	250005	cYIR05	Calib 05	"ratio 1:1"	APB	Calib 05
6	250006	cYIR06	Calib 06	"ratio 1:1"	APB	Calib 06
7	250007	cYIR07	Calib 07	"ratio 1:1"	APB	Calib 07
8	250008	cYIR08	Calib 08	"ratio 1:1"	APB	Calib 08
9	250009	cYIR09	Calib 09	"ratio 1:1"	APB	Calib 09
10	250010	cYIR10	Calib 10	"ratio 1:1"	APB	Calib 10
11	250019	uYIR1	Utility 1	"ratio user defined"	APB	Util
12	250020	uYIR2	Utility 2	"ratio user defined"	APB	Util
13	250011	rYIR1	Ratio 1	"ratio 1:3 low" APB	Ratio 1	Unknown Gene
14	250012	rYIR2	Ratio 2	"ratio 3:1 low" APB	Ratio 2	Unknown Gene
15	250013	rYIR3	Ratio 3	"ratio 1:3 high"	APB	Ratio 3 Unkn
16	250014	rYIR4	Ratio 4	"ratio 3:1 high"	APB	Ratio 4 Unkn
17	13	2515389	APOE	"apolipoprotein E"	GS	Hs.169401
18	37	2375329	BMP5	"bone morphogenetic protein 5"	GS	Hs.1
19	14	1403294		"ESTs" GS	Hs.95612	Unknown Gene
20	38	1384190	ACO2	"aconitase 2, mitochondrial"	GS	Hs.3
21	61	1482116	EML1	"echinoderm microtubule associated protein l		
22	85	1483127	BSN	"bassoon (presynaptic cytomatrix protein)"		
23	74	1330492	PLD1	"phospholipase D1, phophatidylcholine-specif		
24	3	1439929	BRD3	"bromodomain containing 3"	GS	Hs.8
25	208	2472853	HA-1	"minor histocompatibility antigen HA-1" GS		
26	232	2918558	UTY	"ubiquitously transcribed tetratricopeptide		
27	209	1593082	PCDH17	"protocadherin 17"	GS	Hs.106511
28	233	2963871		"" GS	Hs.129764	Unknown Gene
29	256	1290568	R3HDM	"R3H domain (binds single-stranded nucleic a		
30	280	1732346	NEO1	"neogenin homolog 1 (chicken)" GS		Hs.9
31	257	3035874	DLEU2	"deleted in lymphocytic leukemia, 2"	GS	
32	281	1506978	DCAMKL1	"doublecortin and CaM kinase-like 1"	GS	
33	464	3360454	KEO4	"similar to Caenorhabditis elegans protein C		
34	393	2907827	KIAA0321	"KIAA0321 protein"	GS	Hs.8
35	477	1719478	SCN9A	"sodium channel, voltage-gated, type IX, alp		
36	406	1668024		"" GS	Hs.283822	Unknown Gene
37	417	3013646	PPP1R3A	"protein phosphatase 1, regulatory (inhibito		
38	453	2180014	VIPR2	"vasoactive intestinal peptide receptor 2"		
39	430	2189383	SPTB	"spectrin, beta, erythrocytic (includes sphe		
40	454	1860674	CCBL1	"cysteine conjugate-beta lyase; cytoplasmic		
41	746	1634209	ALPI	"alkaline phosphatase, intestinal"	GS	
42	675	1342744	ARHG	"ras homolog gene family, member G (rho G)"		
43	747	1747909	CXCL9	"chemokine (C-X-C motif) ligand 9"	GS	
44	688	1749102	INDO	"indoleamine-pyrrole 2,3 dioxygenase"	GS	
45	699	2190170	GDBR1	"putative glialblastoma cell differentiation		
46	723	1555701	GCS1	"glucosidase I" GS	Hs.83919	Know
47	712	1663080	EIF1A	"eukaryotic translation initiation factor 1A		
48	736	1688943		"EST" GS	Hs.179566	Known Gene
49	949	2903984	GCK	"glucokinase (hexokinase 4, maturity onset d		
50	950	2905885	RAG1	"recombination activating gene 1"	GS	
51	941	3407861	SLC26A4	"solute carrier family 26, member 4"	GS	
52	930	3618382	RGS12	"regulator of G-protein signalling 12"	GS	
53	939	3010959	MYBPH	"myosin binding protein H"	GS	Hs.9
54	940	3614903	KIAA0033	"KIAA0033 protein"	GS	Hs.1
55	955	3744592		"Homo sapiens PP3781 mRNA, complete cds"		
56	944	2961526		"Homo sapiens cDNA FLJ36771 fis, clone ADRGL		
57	1225	1699984	TCFL1	"transcription factor-like 1"	GS	Hs.2
58	1166	1521657	SMARCB1	"SWI/SNF related, matrix associated, actin d		
59	1179	1930711	ITGB4	"integrin, beta 4"	GS	Hs.85266

60	1203	1712592	INSIG1	"insulin induced gene 1"	GS	Hs.5
61	1226	1508336	HBP1	"HMG-box containing protein 1"	GS	Hs.1
62	1155	1905593	PPIG	"peptidyl-prolyl isomerase G (cyclophilin G)		
63	1239	1443061	CNN2	"calponin 2"	GS	Hs.169718 Know
64	1168	2172334	IL6ST	"interleukin 6 signal transducer (gp130, onc		
65	1349	2372824	KIAA0255	"KIAA0255 gene product"	GS	Hs.7
66	1385	2378796	TEKT2	"tektin 2 (testicular)"	GS	Hs.127111
67	1374	2921194	LGALS3	"lectin, galactoside-binding, soluble, 3 (ga		
68	1398	2671453		"" GS Hs.8551 Unknown Gene		
69	1421	1363832	TM9SF1	"transmembrane 9 superfamily member 1"	GS	
70	1350	2955178		"Homo sapiens, clone MGC:19839 IMAGE:4100006		
71	1422	1624865	SNRPD3	"small nuclear ribonucleoprotein D3 polypept		
72	1363	1677133	KIAA0601	"KIAA0601 protein"	GS	Hs.1
73	1580	3174410	RNF4	"ring finger protein 4"	GS	Hs.66394
74	1616	3203789	KIAA0195	"KIAA0195 gene product"	GS	Hs.3
75	1617	1283532	TPBG	"trophoblast glycoprotein"	GS	Hs.8
76	1546	3206210	ATP5A1	"ATP synthase, H+ transporting, mitochondria		
77	1545	2582217	TLR3	"toll-like receptor 3"	GS	Hs.29499
78	1581	3139163	PTGS2	"prostaglandin-endoperoxide synthase 2 (pros		
79	1570	1315115	PYGL	"phosphorylase, glycogen; liver (Hers diseas		
80	1594	2657149		"EST, Highly similar to NRP2_HUMAN Neuropili		
81	1816	1645179	RABIF	"RAB interacting factor"	GS	Hs.9
82	1818	1655365	PCNT2	"pericentrin 2 (kendrin)"	GS	Hs.1
83	1824	1976071	NOC4	"neighbor of COX4"	GS	Hs.173162
84	1849	1813269	CES1	"carboxylesterase 1 (monocyte/macrophage ser		
85	1820	1666268	DNCLI2	"dynein, cytoplasmic, light intermediate pol		
86	1822	1798179		"EST" GS Hs.118021		Known Gene
87	1885	1860355	PHKB	"phosphorylase kinase, beta"	GS	Hs.7
88	1909	1886568	CTCF	"CCCTC-binding factor (zinc finger protein)"		
89	2078	1644303		"ESTs" GS Hs.18081		Unknown Gene
90	2043	1694139	ATP6V1C1	"ATPase, H+ transporting, lysosomal		
91	2020	1702767	PAFAH1B1	"platelet-activating factor acetylhy		
92	2044	1630553	ACE	"angiotensin I converting enzyme (peptidyl-d		
93	2067	1622987	FCGR1A	"Fc fragment of IgG, high affinity Ia, recep		
94	2091	1707220	USP14	"ubiquitin specific protease 14 (tRNA-guanin		
95	2068	1633286	TM4SF4	"transmembrane 4 superfamily member 4"	GS	
96	2092	1910705		"EST" GS Hs.8078 Known Gene		
97	2225	1685883	RB1CC1	"RB1-inducible coiled-coil 1"	GS	Hs.5
98	2249	1658320	HEXA	"hexosaminidase A (alpha polypeptide)"	GS	
99	2238	1556061	ICAM1	"intercellular adhesion molecule 1 (CD54), h		
100	2262	2121653	TNFRSF14	"tumor necrosis factor receptor supe		
101	2273	3217681	ZNF151	"zinc finger protein 151 (pHZ-67)"	GS	
102	2297	3410538	SCA7	"spinocerebellar ataxia 7 (olivopontocerebel		
103	2286	3509885	PDE4C	"phosphodiesterase 4C, cAMP-specific (phosph		
104	2215	1666094	ASNA1	"arsA arsenite transporter, ATP-binding, hom		
105	2409	4017526	SLC6A7	"solute carrier family 6 (neurotransmitter t		
106	2457	2657738	KIAA0220	"KIAA0220 protein"	GS	Hs.1
107	2446	2814416	ZFX	"zinc finger protein, X-linked"	GS	Hs.2
108	2470	2990692		"" GS Hs.290905		Unknown Gene
109	2493	2159205	MPHOSPH1	"M-phase phosphoprotein 1"	GS	
110	2422	3181360	NUP62	"nucleoporin 62kDa"	GS	Hs.9877 Know
111	2494	1724856	FLJ22127	"hypothetical protein FLJ22127"	GS	
112	2423	2841566	BCE-1	"BCE-1 protein" GS	Hs.99824	Know
113	2687	1910527		"EST" GS Hs.13895		Unknown Gene
114	2689	3427294		"ESTs" GS Hs.97042		Unknown Gene
115	2761	2373457		"ESTs" GS Hs.296100		Unknown Gene
116	2690	2514005		"ESTs" GS Hs.214343		Unknown Gene
117	2713	3603182		"ESTs" GS Hs.244461		Unknown Gene
118	2737	1653974		"EST, Weakly similar to 2109260A B cell grow		
119	2714	1233542		"" GS Hs.194359		Unknown Gene

120	2738	2964448		"EST, Highly similar to FYVE-finger-containi			
121	2967	1398814		""	GS	Hs.155489	Unknown Gene
122	2896	1908884		"ESTs"	GS	Hs.132208	Unknown Gene
123	2968	1849552	PRKAR1A	"protein kinase, cAMP-dependent, regulatory,			
124	2897	2691093		"ESTs, Weakly similar to hypothetical protei			
125	2920	1296760		""	GS	Hs.127630	Unknown Gene
126	2944	1349433		"ESTs"	GS	Hs.121070	Unknown Gene
127	2921	2343403		""	GS	Hs.288988	Unknown Gene
128	2957	2594632	RPS27L	"ribosomal protein S27-like"	GS	Hs.1	
129	3139	1467253	ABCA1	"ATP-binding cassette, sub-family A (ABC1),			
130	3163	2175032		""	GS	Hs.26703	Unknown Gene
131	3152	1520041	FLJ10199	"hypothetical protein FLJ10199"	GS		
132	3081	3473302		"ESTs, Highly similar to unknown [Homo sapie			
133	3104	1457140	MAN1B1	"mannosidase, alpha, class 1B, member 1"			
134	3128	1456784	DKFZP566H073	"DKFZP566H073 protein"	GS	Hs.7	
135	3105	1690314		""	GS	Hs.124839	Unknown Gene
136	3129	1490177	FLJ23751	"hypothetical protein FLJ23751"	GS		
137	3297	1805948	TUCAN	"tumor up-regulated CARD-containing antagoni			
138	3321	2618859		""	GS	Hs.337322	Unknown Gene
139	3310	2230024	ARL5	"ADP-ribosylation factor-like 5"	GS		
140	3334	1665871	C1orf19	"chromosome 1 open reading frame 19"	GS		
141	3345	2785292		""	GS	Hs.111680	Unknown Gene
142	3274	1798965	FLJ20093	"hypothetical protein FLJ20093"	GS		
143	3358	1579596		"ESTs, Moderately similar to hypothetical pr			
144	3287	1975219	SLC26A6	"solute carrier family 26, member 6"	GS		
145	3542	1819267		""	GS	Hs.6986	Unknown Gene
146	3544	1829725		""	GS	Hs.278314	Unknown Gene
147	3547	2189366		"Novel human gene mapping to chomosome 1"			
148	3548	1309504	SMC6	"SMC6 protein"	GS	Hs.34497	Know
149	3545	1283020	LOC221955	"KCCR13L"	GS	Hs.131899	
150	3546	2082559	PLSCR3	"phospholipid scramblase 3"	GS	Hs.1	
151	3549	2349407		""	GS	Hs.5888	Unknown Gene
152	3550	1335414		"Homo sapiens, clone IMAGE:4251653, mRNA, pa			
153	3732	2156769		"EST, Moderately similar to PC4259 ferritin			
154	3757	2226159		"ESTs"	GS	Hs.86970	Unknown Gene
155	3829	2350525	ALB	"albumin"	GS	Hs.184411	Know
156	3758	2197506	ZDHHC3	"zinc finger, DHHC domain containing 3"	GS		
157	3781	1485967		""	GS	Hs.127832	Unknown Gene
158	3805	2082211		""	GS	Hs.183765	Unknown Gene
159	3782	1720816	PCDH20	"protocadherin 20"	GS	Hs.132892	
160	3806	1294014		"ESTs"	GS	Hs.293663	Unknown Gene
161	3938	3943727		""	GS	Hs.288038	Unknown Gene
162	3962	2885394		"ESTs, Weakly similar to PRO0478 protein [Ho			
163	3939	4044967	FLJ20535	"hypothetical protein FLJ20535"	GS		
164	3975	3090080	TSP-NY	"testis-specific protein TSP-NY"	GS		
165	3986	4073268		"ESTs"	GS	Hs.133181	Unknown Gene
166	4010	2795249	SLC25A5	"solute carrier family 25 (mitochondrial car			
167	3999	3039249		"Homo sapiens cDNA FLJ14739 fis, clone NT2RP			
168	4023	2652935		""	GS	Hs.74642	Unknown Gene
169	4193	1397294	CROC4	"transcriptional activator of the c-fos prom			
170	4217	2605804	PTPN21	"protein tyrosine phosphatase, non-receptor			
171	4194	2182901	DEFA6	"defensin, alpha 6, Paneth cell-specific"			
172	4218	2827358		"EST"	GS	Hs.24143	Known Gene
173	4146	2794616	PAX6	"paired box gene 6 (aniridia, keratitis)"			
174	4170	1236655	HNRPA1	"heterogeneous nuclear ribonucleoprotein A1"			
175	4147	2867091	FOXJ1	"forkhead box J1"	GS	Hs.93974	
176	4171	4249823	GOLGA4	"golgi autoantigen, golgin subfamily a, 4"			
177	4393	3119252	TOM1	"target of myb1 (chicken)"	GS	Hs.9	
178	4382	2902903	IFITM1	"interferon induced transmembrane protein 1			
179	4384	3942594	PRDX1	"peroxiredoxin 1"	GS	Hs.180909	

180	4408	3246882	ACTR1B	"ARP1 actin-related protein 1 homolog B, cen	
181	4406	4250095	SMARCA5	"SWI/SNF related, matrix associated, actin d	
182	4395	3494714	TJP2	"tight junction protein 2 (zona occludens 2)	
183	4397	4012536	BICD1	"Bicaudal D homolog 1 (Drosophila)"	GS
184	4386	4083705	HLA-DQB1	"major histocompatibility complex, c	
185	4543	254890	PPY	"pancreatic polypeptide"	GS Hs.1
186	4567	340155	INPP5D	"inositol polyphosphate-5-phosphatase, 145kD	
187	4568	809929	ACLY	"ATP citrate lyase"	GS Hs.174140
188	4592	586585	FIBP	"fibroblast growth factor (acidic) intracell	
189	4591	443631	EREG	"epiregulin"	GS Hs.115263 Know
190	4615	2291356	ADAMTS3	"a disintegrin-like and metalloprotease (rep	
191	4628	545165		""	GS Hs.180669 Unknown Gene
192	4557	335707	P5-1	"MHC class I region ORF"	GS Hs.1
193	4846	155892	GUCY1A2	"guanylate cyclase 1, soluble, alpha 2"	GS
194	4870	2054525	ARPC1A	"actin related protein 2/3 complex, subunit	
195	4847	3188155		"ESTs"	GS Hs.227630 Unknown Gene
196	4871	1997302	NSMAF	"neutral sphingomyelinase (N-SMase) activati	
197	4894	3149480	AMHR2	"anti-Mullerian hormone receptor, type II"	
198	4823	1365962	HSPB2	"heat shock 27kDa protein 2"	GS Hs.7
199	4895	2008351	SOX9	"SRY (sex determining region Y)-box 9 (campo	
200	4836	103669	MCM4	"MCM4 minichromosome maintenance deficient 4	
201	5088	2254530	D8S2298E	"reproduction 8"	GS Hs.1
202	5017	1998193	CBFB	"core-binding factor, beta subunit"	GS
203	5018	438438	EPS8	"epidermal growth factor receptor pathway su	
204	7173	G9D1T7 N96785		"A. thaliana DNA for larger subunit of Rubis	
205	5053	1230594	ATRN	"attractin"	GS Hs.194019 Know
206	5089	2009053	ALDH3A2	"aldehyde dehydrogenase 3 family, member A2"	
207	5054	1998342		""	GS Hs.5181 Unknown Gene
208	31127	PolyA50 PolyA50		""	MAF Unknown Gene
209	40073	IMAGp956D011	T73997	BAT2	"HLA-B associated transcript 2" RZPD
210	40193	IMAGp956I011	T65325	TRABID	"TRAF-binding protein domain" RZPD
211	40337	IMAGp956O011	T65570	HERPUD1	"homocysteine-inducible, endoplasmic
212	40026	IMAGp956B021	R38172		"Homo sapiens cDNA FLJ23769 fis, clo
213	40241	IMAGp956K011	T65022	TM4SF11	"transmembrane 4 superfamily member
214	40289	IMAGp956M011	T65542	EGFL6	"EGF-like-domain, multiple 6" RZPD
215	40074	IMAGp956D021	T74051		"" RZPD Unknown Gene
216	40290	IMAGp956M021	R15780	TCF12	"transcription factor 12 (HTF4, heli
217	40509	IMAGp956F052	R38728	KIAA0711	"KIAA0711 gene product" RZPD
218	40605	IMAGp956J052	R37940	RARS	"arginyl-tRNA synthetase" RZPD
219	40438	IMAGp956C062	R39356	TP53	"tumor protein p53 (Li-Fraumeni synd
220	40582	IMAGp956I062	R39439	BACE	"beta-site APP-cleaving enzyme" RZPD
221	40653	IMAGp956L052	R37604	HGF	"hepatocyte growth factor (hepapoiet
222	40725	IMAGp956O052	R39241	RCOR	"REST corepressor" RZPD Hs.7
223	40702	IMAGp956N062	T78930	RANBP2	"RAN binding protein 2" RZPD Hs.1
224	40439	IMAGp956C072	R38705	EIF2B3	"eukaryotic translation initiation f
225	40879	IMAGp956E153	T81077	FOXO3A	"forkhead box O3A" RZPD Hs.3
226	41000	IMAGp956J163	T81576	GPHN	"gephyrin" RZPD Hs.13405
227	41122	IMAGp956O183	T83179	HSPC152	"hypothetical protein HSPC152" RZPD
228	40931	IMAGp956G193	T81384	GRHPR	"glyoxylate reductase/hydroxypyruvat
229	40857	IMAGp956D173	T70303	NDUFB2	"NADH dehydrogenase (ubiquinone) 1 b
230	40953	IMAGp956H173	T70293	DAZAP2	"DAZ associated protein 2" RZPD
231	41075	IMAGp956M193	T77816	CCL2	"chemokine (C-C motif) ligand 2"
232	40836	IMAGp956C203	T72639	HMGCS2	"3-hydroxy-3-methylglutaryl-Coenzyme
233	41716	IMAGp956H125	T82202	TM7SF3	"seven transmembrane protein TM7SF3"
234	41788	IMAGp956K125	T78093	BTAF1	"BTAF1 RNA polymerase II, B-TFIID tr
235	41790	IMAGp956K145	T78405		"ESTs" RZPD Hs.396135 Unkn
236	41910	IMAGp956P145	T87345	FLJ10035	"hypothetical protein FLJ100
237	41765	IMAGp956J135	T86478	FLJ20300	"hypothetical protein FLJ203
238	41550	IMAGp956A145	T86884	MGC12760	"hypothetical protein MGC127
239	41623	IMAGp956D155	T78859	MGC4342	"hypothetical protein MGC4342" RZPD

240	41791	IMAGp956K155	T77062	NCOA3	"nuclear receptor coactivator 3"
241	42668	IMAGp956P047	R01713	KIAA0101	"KIAA0101 gene product" RZPD
242	42381	IMAGp956D057	T99236	JUNB	"jun B proto-oncogene" RZPD Hs.1
243	42430	IMAGp956F067	T98007	HSRTSBETA	"rTS beta protein" RZPD
244	42407	IMAGp956E077	T95099		"Homo sapiens, RNA helicase-related
245	42477	IMAGp956H057	T98652	GBE1	"glucan (1,4-alpha-), branching enzy
246	42669	IMAGp956P057	T98952	DSG1	"desmoglein 1" RZPD Hs.2633 Know
247	42599	IMAGp956M077	T95334	XPO4	"likely ortholog of mouse exportin 4
248	42480	IMAGp956H087	R00496	GALNT2	"UDP-N-acetyl-alpha-D-galactosamine:
249	43121	IMAGp956C019	R00942		"" RZPD Unknown Gene
250	43361	IMAGp956M019	R06106	SLC38A3	"solute carrier family 38, member 3"
251	43338	IMAGp956L029	R08182	DUSP16	"dual specificity phosphatase 16"
252	43434	IMAGp956P029	R06596		"" RZPD Unknown Gene
253	43146	IMAGp956D029	R06414	OGN	"osteoglycin (osteoinductive factor,
254	43290	IMAGp956J029	R07172	HCNGP	"transcriptional regulator protein"
255	43219	IMAGp956G039	R00915	VPS4B	"vacuolar protein sorting 4B (yeast)
256	43267	IMAGp956I039	R05792	KIAA0781	"KIAA0781 protein" RZPD
257	43807	IMAGp956O1510	R09604	EIF3S10	"eukaryotic translation initiation f
258	43544	IMAGp956D1610	R10351	LCP	"host cell factor homolog" RZPD
259	43497	IMAGp956B1710	R12690	SOX13	"SRY (sex determining region Y)-box
260	43713	IMAGp956K1710	R09416	CD36	"CD36 antigen (collagen type I recep
261	43592	IMAGp956F1610	R16478	RBSK	"ribokinase" RZPD Hs.11916
262	43784	IMAGp956N1610	R10907	PCDHGC3	"protocadherin gamma subfamily C, 3"
263	43498	IMAGp956B1810	R10328	OBTP	"over-expressed breast tumor protein
264	43570	IMAGp956E1810	R08643		"Homo sapiens, clone IMAGE:3882977,
265	44005	IMAGp956G2111	R12261	ANLN	"anillin, actin binding protein (scr
266	44149	IMAGp956M2111	R16601	HSPC065	"HSPC065 protein" RZPD Hs.1
267	44054	IMAGp956I2211	R11469	IDH3G	"isocitrate dehydrogenase 3 (NAD+) g
268	44198	IMAGp956O2211	R11625		"ESTs" RZPD Hs.374869 Unkn
269	43886	IMAGp956B2211	R12045	PXMP4	"peroxisomal membrane protein 4, 24k
270	44006	IMAGp956G2211	R17009		"ESTs" RZPD Hs.348862 Unkn
271	43863	IMAGp956A2311	R11663	BIRC6	"baculoviral IAP repeat-containing 6
272	43935	IMAGp956D2311	R12044	KIF13B	"kinesin family member 13B" RZPD
273	44510	IMAGp956L2212	R15384	RANBP7	"RAN binding protein 7" RZPD Hs.3
274	44271	IMAGp956B2312	R13383	GLRB	"glycine receptor, beta" RZPD
275	44296	IMAGp956C2412	R13764	CENTB5	"centaurin, beta 5" RZPD Hs.2
276	44368	IMAGp956F2412	R13136	HAPIP	"huntingtin-associated protein inter
277	44391	IMAGp956G2312	R12850	TM6SF1	"transmembrane 6 superfamily member
278	44439	IMAGp956I2312	R38426	CTNNBL1	"catenin, beta like 1" RZPD Hs.1
279	44440	IMAGp956I2412	R37165	CNOT7	"CCR4-NOT transcription complex, sub
280	44536	IMAGp956M2412	R13291	GEM	"GTP binding protein overexpressed i
281	45353	IMAGp956P0114	R20300	HSPC194	"hypothetical protein HSPC194" RZPD
282	45042	IMAGp956C0214	R20139	NFE2L3	"nuclear factor (erythroid-derived 2
283	45282	IMAGp956M0214	R20303	NAT5	"N-acetyltransferase 5 (ARD1 homolog
284	44995	IMAGp956A0314	R18743	SEPP1	"selenoprotein P, plasma, 1" RZPD
285	45090	IMAGp956E0214	R36753	KIAA1165	"likely ortholog of mouse Ne
286	45162	IMAGp956H0214	R24861	ZNF289	"zinc finger protein 289, ID1 regula
287	45139	IMAGp956G0314	R20968	SHANK2	"SH3 and multiple ankyrin repeat dom
288	45187	IMAGp956I0314	R17278	MALT1	"mucosa associated lymphoid tissue l
289	45618	IMAGp956K0215	R19732		"" RZPD Unknown Gene
290	45690	IMAGp956N0215	R25797	SLC22A6	"solute carrier family 22 (organic a
291	45691	IMAGp956N0315	R45448	ZNF261	"zinc finger protein 261" RZPD
292	45452	IMAGp956D0415	R14662	NJMU-R1	"protein kinase Njmu-R1" RZPD
293	45427	IMAGp956C0315	R19651	LAMR1	"laminin receptor 1 (ribosomal prote
294	45475	IMAGp956E0315	R19959	GABRB3	"gamma-aminobutyric acid (GABA) A re
295	45548	IMAGp956H0415	R34891	ABCC5	"ATP-binding cassette, sub-family C
296	45596	IMAGp956J0415	R20997	PF20	"PF20" RZPD Hs.6783 Known Gene
297	45977	IMAGp956J0116	R54397	HTATIP	"HIV-1 Tat interactive protein, 60kD
298	46121	IMAGp956P0116	R51177	EIF2AK3	"eukaryotic translation initiation f
299	45787	IMAGp956B0316	R50762	MAEA	"macrophage erythroblast attacher"

300	45907	IMAGp956G0316	R26483	FLJ10648	"function unknown protein 1"
301	45954	IMAGp956I0216	R26621	FAIM2	"Fas apoptotic inhibitory molecule 2
302	46122	IMAGp956P0216	R52461	KIAA0781	"KIAA0781 protein" RZPD
303	46123	IMAGp956P0316	R51384	PTDSS2	"phosphatidylserine synthase 2" RZPD
304	45788	IMAGp956B0416	R36025	SDBCAG84	"serologically defined breas
305	46337	IMAGp956I0117	R54061	HR	"hairless" RZPD Hs.272367
306	46146	IMAGp956A0217	R52093	FLJ13263	"hypothetical protein FLJ132
307	46267	IMAGp956F0317	R56221	DUSP6	"dual specificity phosphatase 6"
308	46339	IMAGp956I0317	R54792	PCNP	"PEST-containing nuclear protein"
309	46314	IMAGp956H0217	R56229	SLC25A5	"solute carrier family 25 (mitochond
310	46458	IMAGp956N0217	R60892	ADCY8	"adenylate cyclase 8 (brain)" RZPD
311	46411	IMAGp956L0317	R56085		"" RZPD Unknown Gene
312	46459	IMAGp956N0317	R58957	DKFZP547E1010	"DKFZP547E1010 protein" RZPD
313	46723	IMAGp956I0318	R60299	EXTL3	"exostoses (multiple)-like 3" RZPD
314	46795	IMAGp956L0318	R60049	CELSR2	"cadherin, EGF LAG seven-pass G-type
315	46652	IMAGp956F0418	R21534	FBLN1	"fibulin 1" RZPD Hs.79732
316	46700	IMAGp956H0418	R61395	ECEL1	"endothelin converting enzyme-like 1
317	46532	IMAGp956A0418	R61163	COL4A5	"collagen, type IV, alpha 5 (Alport
318	46604	IMAGp956D0418	R61382	ASNA1	"arsA arsenite transporter, ATP-bind
319	46748	IMAGp956J0418	R21919	PELI1	"pellino homolog 1 (Drosophila)"
320	46844	IMAGp956N0418	R21727	DYSF	"dysferlin, limb girdle muscular dys
321	47027	IMAGp956E1919	R60561	MRPS11	"mitochondrial ribosomal protein S11
322	46932	IMAGp956A2019	R26460	BNIP3L	"BCL2/adenovirus E1B 19kDa interacti
323	46933	IMAGp956A2119	R22509	C21orf59	"chromosome 21 open reading
324	47101	IMAGp956H2119	R25877	FLJ10511	"hypothetical protein FLJ105
325	47124	IMAGp956I2019	R23942	ALPP	"alkaline phosphatase, placental (Re
326	47196	IMAGp956L2019	R31779	CTNNBIP1	"catenin, beta interacting p
327	47293	IMAGp956P2119	R25244	CLPX	"ClpX caseinolytic protease X homolo
328	46958	IMAGp956B2219	R31932	FLJ13964	"hypothetical protein FLJ139
329	47946	IMAGp956L0221	R70497	XRCC3	"X-ray repair complementing defectiv
330	48018	IMAGp956O0221	R68683	HSD11B2	"hydroxysteroid (11-beta) dehydrogen
331	47995	IMAGp956N0321	R64113	CPD	"carboxypeptidase D" RZPD Hs.5
332	47780	IMAGp956E0421	R62613	TRIM37	"tripartite motif-containing 37"
333	47875	IMAGp956I0321	R62184	NCOA3	"nuclear receptor coactivator 3"
334	47923	IMAGp956K0321	R36149	KIAA0970	"KIAA0970 protein" RZPD
335	47900	IMAGp956J0421	R66885	FHL2	"four and a half LIM domains 2" RZPD
336	47996	IMAGp956N0421	R68777	YES1	"v-yes-1 Yamaguchi sarcoma viral onc
337	48287	IMAGp956J0722	R78630	C1orf13	"chromosome 1 open reading frame 13"
338	48383	IMAGp956N0722	R77909	GLCE	"likely homolog of mouse glucuronyl
339	48240	IMAGp956H0822	R79193	CLN2	"ceroid-lipofuscinosis, neuronal 2,
340	48312	IMAGp956K0822	R74174		"ESTs" RZPD Hs.381521 Unkn
341	48431	IMAGp956P0722	R71140	P5	"protein disulfide isomerase-related
342	48192	IMAGp956F0822	R79157	MYO5C	"myosin 5C" RZPD Hs.111782
343	48169	IMAGp956E0922	R72966		"ESTs" RZPD Hs.355920 Unkn
344	48241	IMAGp956H0922	R77756	CRSP3	"cofactor required for Sp1 transcrip
345	48677	IMAGp956J1323	R81979	SYPL	"synaptophysin-like protein" RZPD
346	48510	IMAGp956C1423	R82824	DKFZP761C169	"hypothetical protein DKFZp7
347	48559	IMAGp956E1523	R80602	FLJ14281	"hypothetical protein FLJ142
348	48631	IMAGp956H1523	H01524	GALC	"galactosylceramidase (Krabbe diseas
349	48558	IMAGp956E1423	R81498	FLJ20445	"hypothetical protein FLJ204
350	48654	IMAGp956I1423	R82480	HIP-55	"src homology 3 domain-containing pr
351	48727	IMAGp956L1523	H04399	HSPCB	"heat shock 90kDa protein 1, beta"
352	48823	IMAGp956P1523	H12337		"" RZPD Unknown Gene
353	48849	IMAGp956A1724	H00686	TBX3	"T-box 3 (ulnar mammary syndrome)"
354	48921	IMAGp956D1724	R67654	MGC11279	"hypothetical protein MGC112
355	49161	IMAGp956N1724	R48263	TIAF1	"TGFB1-induced anti-apoptotic factor
356	49209	IMAGp956P1724	R50087	GREB1	"GREB1 protein" RZPD Hs.193914
357	48969	IMAGp956F1724	R50169	FGF11	"fibroblast growth factor 11" RZPD
358	49089	IMAGp956K1724	H03832	C1orf22	"chromosome 1 open reading frame 22"
359	48898	IMAGp956C1824	H03267	TFPI2	"tissue factor pathway inhibitor 2"

360	48946	IMAGp956E1824	R46507	UBE2V1	"ubiquitin-conjugating enzyme E2 var
361	49305	IMAGp956D1725	H21675	HCA127	"hepatocellular carcinoma-associated
362	49401	IMAGp956H1725	H25003	IF	"I factor (complement)" RZPD Hs.3
363	49378	IMAGp956G1825	R70251	LOC56834	"chromosome 11 hypothetical
364	49450	IMAGp956J1825	H15742	OSBPL8	"oxysterol binding protein-like 8"
365	49497	IMAGp956L1725	R70223	NUP107	"nuclear pore complex protein" RZPD
366	49258	IMAGp956B1825	H21724	USP24	"ubiquitin specific protease 24"
367	49498	IMAGp956L1825	H27331	MGC5338	"hypothetical protein MGC5338" RZPD
368	49546	IMAGp956N1825	H21764	APOL2	"apolipoprotein L, 2" RZPD Hs.2
369	49908	IMAGp956M2026	H27244	HSPC166	"HSPC166 protein" RZPD Hs.2
370	49621	IMAGp956A2126	H06473	PDCD6IP	"programmed cell death 6 interacting
371	49861	IMAGp956K2126	H26078	PRKCN	"protein kinase C, nu" RZPD Hs.1
372	49933	IMAGp956N2126	H05383	LGI1	"leucine-rich, glioma inactivated 1"
373	49669	IMAGp956C2126	H06194	MGC2752	"hypothetical protein MGC2752" RZPD
374	49789	IMAGp956H2126	H06531	DNAJA1	"DnaJ (Hsp40) homolog, subfamily A,
375	49622	IMAGp956A2226	H05223	C20orf39	"chromosome 20 open reading
376	49718	IMAGp956E2226	H05615	LOC51097	"CGI-49 protein" RZPD
377	50197	IMAGp956I2127	H09512	NIPSNAP1	"nipsnap homolog 1 (C. elega
378	50293	IMAGp956M2127	H08409	SEC61A2	"likely ortholog of mouse SEC61, alp
379	50150	IMAGp956G2227	H09626	PRKRA	"protein kinase, interferon-inducibl
380	50222	IMAGp956J2227	H09996	MMP16	"matrix metalloproteinase 16 (membra
381	50365	IMAGp956P2127	H11275	ZNF264	"zinc finger protein 264" RZPD
382	50030	IMAGp956B2227	H11901	ATPAF1	"ATP synthase mitochondrial F1 compl
383	50270	IMAGp956L2227	H12229	FLJ20080	"hypothetical protein FLJ200
384	50342	IMAGp956O2227	H09810	CAMTA1	"calmodulin binding transcription ac
385	46753	IMAGp956J0918	R60028	MAP3K7IP1	"mitogen-activated protein k
386	46562	IMAGp956B1018	R20806	SOAT1	"sterol O-acyltransferase (acyl-Coen
387	46659	IMAGp956F1118	R59938	RRM1	"ribonucleotide reductase M1 polypep
388	46899	IMAGp956P1118	R60245	RAB31	"RAB31, member RAS oncogene family"
389	46730	IMAGp956I1018	R60014	DKFZP761N09121	"hypothetical protein DKFZp7
390	46539	IMAGp956A1118	R60283	GOT1	"glutamic-oxaloacetic transaminase 1
391	46612	IMAGp956D1218	R20802	EP400	"E1A binding protein p400" RZPD
392	46756	IMAGp956J1218	R21182	RAB11A	"RAB11A, member RAS oncogene family"
393	51449	IMAGp956N0130	R87514	CDKN1A	"cyclin-dependent kinase inhibitor 1
394	51234	IMAGp956E0230	H18981	CCNG2	"cyclin G2" RZPD Hs.79069
395	51379	IMAGp956K0330	H29604		"Homo sapiens cDNA FLJ38264 fis, clo
396	51188	IMAGp956C0430	R89046	ABCB8	"ATP-binding cassette, sub-family B
397	51235	IMAGp956E0330	H29581	STMN4	"stathmin-like 4" RZPD Hs.3
398	51283	IMAGp956G0330	R87572	ERG	"v-ets erythroblastosis virus E26 on
399	51236	IMAGp956E0430	H18233	HEY1	"hairy/enhancer-of-split related wit
400	51308	IMAGp956H0430	H20145	ACATN	"acetyl-Coenzyme A transporter" RZPD
401	51648	IMAGp956F0831	R87763	ICAM5	"intercellular adhesion molecule 5,
402	51529	IMAGp956A0931	H46425	PURA	"purine-rich element binding protein
403	51769	IMAGp956K0931	H41801	MMP25	"matrix metalloproteinase 25" RZPD
404	51889	IMAGp956P0931	H50940	TNIP1	"TNFAIP3 interacting protein 1" RZPD
405	51577	IMAGp956C0931	H44908	C21orf51	"chromosome 21 open reading
406	51721	IMAGp956I0931	H46199	DKFZP586J0619	"DKFZP586J0619 protein" RZPD
407	51578	IMAGp956C1031	H45298	FLJ20421	"hypothetical protein FLJ204
408	51722	IMAGp956I1031	H42184	POLK	"polymerase (DNA directed) kappa"
409	52481	IMAGp956I0133	H47585	PVALB	"parvalbumin" RZPD Hs.295449
410	52314	IMAGp956B0233	R88729	REPS2	"RALBP1 associated Eps domain contai
411	52435	IMAGp956G0333	H47327	TACC1	"transforming, acidic coiled-coil co
412	52555	IMAGp956L0333	R88882	CPVL	"carboxypeptidase, vitellogenic-like
413	52554	IMAGp956L0233	R91402	MDS032	"uncharacterized hematopoietic stem/
414	52291	IMAGp956A0333	R84308	CEACAM1	"carcinoembryonic antigen-related ce
415	52627	IMAGp956O0333	H50978	ABCD4	"ATP-binding cassette, sub-family D
416	52292	IMAGp956A0433	R83300	TAF1	"TAF1 RNA polymerase II, TATA box bi
417	53081	IMAGp956B0135	H48533	BIRC3	"baculoviral IAP repeat-containing 3
418	53321	IMAGp956L0135	H48220	RASA1	"RAS p21 protein activator (GTPase a
419	53394	IMAGp956O0235	R98995	FLJ20417	"hypothetical protein FLJ204

420	53323	IMAGp956L0335	H54023	LILRB2	"leukocyte immunoglobulin-like recep
421	53154	IMAGp956E0235	R98627	CASK	"calcium/calmodulin-dependent serine
422	53202	IMAGp956G0235	R98887	KIAA1450	"KIAA1450 protein" RZPD
423	53108	IMAGp956C0435	R98471	H2BFQ	"H2B histone family, member Q" RZPD
424	53324	IMAGp956L0435	R99725	""	RZPD Unknown Gene
425	53828	IMAGp956A0437	H62563	FLJ13409	"hypothetical protein FLJ134
426	54044	IMAGp956J0437	H65325	HNRPDL	"heterogeneous nuclear ribonucleopro
427	53949	IMAGp956F0537	H63689	FLJ20003	"hypothetical protein FLJ200
428	53902	IMAGp956D0637	H64124	CNOT2	"CCR4-NOT transcription complex, sub
429	54116	IMAGp956M0437	H63088	ZDHHC3	"zinc finger, DHHC domain containing
430	53901	IMAGp956D0537	H65328	RAB11B	"RAB11B, member RAS oncogene family"
431	54118	IMAGp956M0637	H63072	PLAC1	"placenta-specific 1" RZPD Hs.1
432	54166	IMAGp956O0637	H60454	MLLT4	"myeloid/lymphoid or mixed-lineage l
433	54883	IMAGp956M0339	H84574	HSD17B8	"hydroxysteroid (17-beta) dehydrogen
434	54668	IMAGp956D0439	H79250	EIF2S1	"eukaryotic translation initiation f
435	54958	IMAGp956P0639	H78741	PTDSR	"phosphatidylserine receptor" RZPD
436	54671	IMAGp956D0739	H85019	KPNB1	"karyopherin (importin) beta 1" RZPD
437	54813	IMAGp956J0539	H86518	ARR3	"arrestin 3, retinal (X-arrestin)"
438	54862	IMAGp956L0639	H78829	FLJ11021	"hypothetical protein FLJ110
439	54767	IMAGp956H0739	H85996	GUCA1A	"guanylate cyclase activator 1A (ret
440	54959	IMAGp956P0739	H86554	CP	"ceruloplasmin (ferroxidase)" RZPD
441	55461	IMAGp956E0541	H67789	FLJ11193	"hypothetical protein FLJ111
442	55629	IMAGp956L0541	H82423	SERPINA5	"serine (or cysteine) protei
443	55582	IMAGp956J0641	H82409	""	RZPD Unknown Gene
444	55702	IMAGp956O0641	H61209	FLJ20195	"hypothetical protein FLJ201
445	55725	IMAGp956P0541	AF074392	NSG-X	"brain and nasopharyngeal ca
446	55438	IMAGp956D0641	H81908	BSG	"basigin (OK blood group)" RZPD
447	55391	IMAGp956B0741	H90310	ZIN	"zinedin" RZPD Hs.108665
448	55368	IMAGp956A0841	H61374	FANCC	"Fanconi anemia, complementation gro
449	56302	IMAGp956H0643	N58398	FLJ12057	"hypothetical protein FLJ120
450	56446	IMAGp956N0643	N58114	E2F4	"E2F transcription factor 4, p107/p1
451	56257	IMAGp956F0943	N57657	FLJ10581	"putative RNA methyltransfer
452	56234	IMAGp956E1043	N57964	CCR6	"chemokine (C-C motif) receptor 6"
453	56184	IMAGp956C0843	N55503	MRPL16	"mitochondrial ribosomal protein L16
454	56376	IMAGp956K0843	N55496	LOC51252	"hypothetical protein LOC512
455	56378	IMAGp956K1043	N52496	BTG3	"BTG family, member 3" RZPD Hs.7
456	56498	IMAGp956P1043	N53133	STRBP	"spermatid perinuclear RNA binding p
457	57045	IMAGp956G0545	N56660	LATS2	"LATS, large tumor suppressor, homol
458	57213	IMAGp956N0545	H99084	GUK1	"guanylate kinase 1" RZPD Hs.3
459	57166	IMAGp956L0645	N20049	KIAA0802	"KIAA0802 protein" RZPD
460	57071	IMAGp956H0745	N20630	GIOT-3	"GIOT-3 for gonadotropin inducible t
461	56974	IMAGp956D0645	H99736	CHD1	"chromodomain helicase DNA binding p
462	57094	IMAGp956I0645	H98694	SMG1	"PI-3-kinase-related kinase SMG-1"
463	57239	IMAGp956O0745	N32787	DKFZP564G092	"DKFZP564G092 protein" RZPD
464	57216	IMAGp956N0845	N21375	ARF6	"ADP-ribosylation factor 6" RZPD
465	57566	IMAGp956L2246	N27421	OAZ1	"ornithine decarboxylase antizyme 1"
466	57447	IMAGp956G2346	N21636	PBP	"prostatic binding protein" RZPD
467	57663	IMAGp956P2346	N24045	CCNDBP1	"cyclin D-type binding-protein 1"
468	57472	IMAGp956H2446	N23400	IFRG28	"28kD interferon responsive protein"
469	57495	IMAGp956I2346	N23213	HIVEP1	"human immunodeficiency virus type I
470	57567	IMAGp956L2346	N24042	DDAH1	"dimethylarginine dimethylaminohydro
471	57592	IMAGp956M2446	N23036	FLJ12890	"hypothetical protein FLJ128
472	57785	IMAGp956F0147	N36863	SERP1	"stress-associated endoplasmic retic
473	58463	IMAGp956B0749	N51702	FLJ21128	"hypothetical protein FLJ211
474	58560	IMAGp956F0849	N45644	IPP	"intracisternal A particle-promoted
475	58538	IMAGp956E1049	N46838		"ESTs, Moderately similar to 2211404
476	58587	IMAGp956G1149	N47623	SP329	"hypothetical protein SP329" RZPD
477	58680	IMAGp956K0849	N45714	HMGN4	"high mobility group nucleosomal bin
478	58465	IMAGp956B0949	N51703	NLGN3	"neuroligin 3" RZPD Hs.47320
479	58492	IMAGp956C1249	N47381	PEPP3	"phosphoinositol 3-phosphate-binding

480	58588	IMAGp956G1249	N46860	MIR16	"membrane interacting protein of RGS
481	59481	IMAGp956L1751	N59179	NEK1	"NIMA (never in mitosis gene a)-rela
482	59529	IMAGp956N1751	N62252	HCC8	"tumor antigen SLP-8p" RZPD Hs.4
483	59578	IMAGp956P1851	N59292	WDR7	"WD repeat domain 7" RZPD Hs.1
484	59363	IMAGp956G1951	N53539	RECQL5	"RecQ protein-like 5" RZPD Hs.3
485	59338	IMAGp956F1851	N59270	IL15	"interleukin 15" RZPD Hs.1
486	59482	IMAGp956L1851	N59279	ZNF3	"zinc finger protein 3 (A8-51)" RZPD
487	59579	IMAGp956P1951	N59818	ZFP318	"endocrine regulator" RZPD Hs.6
488	59460	IMAGp956K2051	N50632	LRDD	"leucine-rich and death domain conta
489	60345	IMAGp956P1753	N64025	FLJ22637	"hypothetical protein FLJ226
490	60034	IMAGp956C1853	N64862	FYB	"FYN binding protein (FYB-120/130)"
491	60251	IMAGp956L1953	N63855	AF020591	"zinc finger protein" RZPD
492	60229	IMAGp956K2153	N66818	MSI2	"musashi homolog 2 (Drosophila)"
493	60178	IMAGp956I1853	N63852	UNG	"uracil-DNA glycosylase" RZPD
494	59987	IMAGp956A1953	N68496	FBXL4	"F-box and leucine-rich repeat prote
495	60277	IMAGp956M2153	N63843	IPP	"intracisternal A particle-promoted
496	60086	IMAGp956E2253	N63903	KIAA1322	"KIAA1322 protein" RZPD
497	60968	IMAGp956J1655	N79582	CPA3	"carboxypeptidase A3 (mast cell)"
498	61064	IMAGp956N1655	N77955	HSPC128	"HSPC128 protein" RZPD Hs.9
499	60922	IMAGp956H1855	N80276	SOX9	"SRY (sex determining region Y)-box
500	60994	IMAGp956K1855	N74673	RARG	"retinoic acid receptor, gamma" RZPD
501	60897	IMAGp956G1755	N74086	FLJ20280	"hypothetical protein FLJ202
502	61113	IMAGp956P1755	N75074	APEH	"N-acylaminoacyl-peptide hydrolase"
503	61114	IMAGp956P1855	N79609	CKN1	"Cockayne syndrome 1 (classical)"
504	60995	IMAGp956K1955	N70579	DKFZP566F084	"DKFZP566F084 protein" RZPD
505	61722	IMAGp956J0257	W32508	STK3	"serine/threonine kinase 3 (STE20 ho
506	61866	IMAGp956P0257	W37287	RAB18	"RAB18, member RAS oncogene family"
507	61867	IMAGp956P0357	W37098	FCMD	"Fukuyama type congenital muscular d
508	61580	IMAGp956D0457	W32162	ZFP161	"zinc finger protein 161 homolog (mo
509	61555	IMAGp956C0357	N92269	SNRPA1	"small nuclear ribonucleoprotein pol
510	61795	IMAGp956M0357	N93925	SCAP2	"src family associated phosphoprotei
511	61652	IMAGp956G0457	W04573	GTF3C2	"general transcription factor IIIC,
512	61748	IMAGp956K0457	N99161	COASTER	"coactivator for steroid receptors"
513	62042	IMAGp956G1058	W47394	SSR1	"signal sequence receptor, alpha (tr
514	62162	IMAGp956L1058	W48574	RASSF1	"Ras association (RalGDS/AF-6) domai
515	62067	IMAGp956H1158	W02718	MLC1SA	"myosin light chain 1 slow a" RZPD
516	62139	IMAGp956K1158	W37506	ASML3B	"acid sphingomyelinase-like phosphod
517	62258	IMAGp956P1058	W46985	MMP24	"matrix metalloproteinase 24 (membra
518	61995	IMAGp956E1158	W42775	DIM1	"similar to S. pombe dim1+" RZPD
519	62235	IMAGp956O1158	AA037412	FLJ10829	"dudulin 2" RZPD
520	61948	IMAGp956C1258	W47118	FLJ12619	"hypothetical protein FLJ126
521	62505	IMAGp956J1759	W67895	SERF1A	"small EDRK-rich factor 1A (telomeri
522	62625	IMAGp956O1759	W02127	ZNF85	"zinc finger protein 85 (HPF4, HTF1)
523	62411	IMAGp956F1959	W69183	PLA2G4B	"phospholipase A2, group IVB (cytoso
524	62292	IMAGp956A2059	W57945	ICOS	"inducible T-cell co-stimulator"
525	62578	IMAGp956M1859	W60806	ALTE	"Ac-like transposable element" RZPD
526	62650	IMAGp956P1859	W67505	DNAJB11	"DnaJ (Hsp40) homolog, subfamily B,
527	62436	IMAGp956G2059	W58281	TERA	"TERA protein" RZPD Hs.180780
528	62317	IMAGp956B2159	W60626	AK2	"adenylate kinase 2" RZPD Hs.1
529	63096	IMAGp956C0861	W96216	NICE-1	"NICE-1 protein" RZPD Hs.1
530	63384	IMAGp956O0861	W93818	CDC2L5	"cell division cycle 2-like 5 (choli
531	63314	IMAGp956L1061	W94861	MDS033	"uncharacterized hematopoietic stem/
532	63410	IMAGp956P1061	AA010313	ADRA2C	"adrenergic, alpha-2C-, rece
533	63217	IMAGp956H0961	W95803	SLC5A1	"solute carrier family 5 (sodium/glu
534	63218	IMAGp956H1061	W95873	MYOZ2	"myozenin 2" RZPD Hs.381047
535	63195	IMAGp956G1161	W93568	KIAA0008	"Drosophila discs large-1 tu
536	63363	IMAGp956N1161	W94613	KIAA1345	"KIAA1345 protein" RZPD
537	63914	IMAGp956E1063	AA022883	RAB6IP1	"RAB6 interacting protein 1"
538	64034	IMAGp956J1063	AA032221	STEAP	"six transmembrane epithelia
539	64108	IMAGp956M1263	AA024822	LCHN	"LCHN protein" RZPD Hs.1

540	63869	IMAGp956C1363	AA019197		GPR75	"G protein-coupled receptor
541	64130	IMAGp956N1063	AA027160		GSPT2	"G1 to S phase transition 2"
542	64012	IMAGp956I1263	AA024857		RPA40	"RNA polymerase I subunit"
543	64013	IMAGp956I1363	AA018856		PDE6G	"phosphodiesterase 6G, cGMP-
544	63870	IMAGp956C1463	AA009454		MYBBP1A	"MYB binding protein (P160)
545	65086	IMAGp956F0666	W88708	BIN3	"bridging integrator 3" RZPD	Hs.6
546	65302	IMAGp956O0666	W88654	RRM2	"ribonucleotide reductase M2 polypep	
547	65304	IMAGp956O0866	W85898	PSTPIP1	"proline-serine-threonine phosphatas	
548	65041	IMAGp956D0966	W88753	PMPCB	"peptidase (mitochondrial processing	
549	65327	IMAGp956P0766	W92805	APBB2	"amyloid beta (A4) precursor protein	
550	65088	IMAGp956F0866	W90174	CSNK1G1	"casein kinase 1, gamma 1"	RZPD
551	65042	IMAGp956D1066	W90118	SLC1A5	"solute carrier family 1 (neutral am	
552	65091	IMAGp956F1166	W88752	MRPL27	"mitochondrial ribosomal protein L27	
553	66302	IMAGp956H2269	AA033573		GARS	"glycyl-tRNA synthetase"
554	66470	IMAGp956O2269	AA029750		PSME2	"proteasome (prosome, macrop
555	66593	IMAGp956E0170	AA034918		KIAA1028	"KIAA1028 protein"
556	66546	IMAGp956C0270	AA040870		DKFZP586I2223	"intermediate filame
557	66256	IMAGp956F2469	AA033640		RE2	"G-protein coupled receptor"
558	66496	IMAGp956P2469	AA033779		UTX	"ubiquitously transcribed te
559	66666	IMAGp956H0270	AA045176		RAP140	"KIAA1105 protein" RZPD
560	66810	IMAGp956N0270	AA045066		APG12L	"APG12 autophagy 12-like (S.
561	67007	IMAGp956F0771	AA128018		FOXL2	"forkhead box L2" RZPD
562	67200	IMAGp956N0871	AA156859		KIAA1018	"KIAA1018 protein"
563	66962	IMAGp956D1071	AA131584		DKFZP564O0463	"DKFZP564O0463 prote
564	67011	IMAGp956F1171	AA115734		NDUFC2	"NADH dehydrogenase (ubiquin
565	67033	IMAGp956G0971	AA099369		CUL5	"cullin 5" RZPD Hs.1
566	67129	IMAGp956K0971	AA047109		ITGBL1	"integrin, beta-like 1 (with
567	66916	IMAGp956B1271	AA126928		IFI16	"interferon, gamma-inducible
568	67084	IMAGp956I1271	AA150292		FLJ12436	"hypothetical protei
569	67744	IMAGp956D2473	AA291675		WFDC2	"WAP four-disulfide core dom
570	67960	IMAGp956M2473	AA278138		FLJ23375	"hypothetical protei
571	68297	IMAGp956L0174	AA292745		SIRT6	"sirtuin silent mating type
572	68058	IMAGp956B0274	AA399268		C21orf80	"chromosome 21 open
573	68081	IMAGp956C0174	AA283640		ZNF337	"zinc finger protein 337"
574	68177	IMAGp956G0174	AA235548		PNKP	"polynucleotide kinase 3'-ph
575	68154	IMAGp956F0274	AA398238		EYA1	"eyes absent homolog 1 (Dros
576	68226	IMAGp956I0274	AA398211		LZTFL1	"leucine zipper transcriptio
577			""		Unknown Gene	
578	68954	IMAGp956G1076	AA393956		ZFP	"zinc finger protein" RZPD
579	69028	IMAGp956J1276	AA421088		BPI	"bactericidal/permeability-i
580	68861	IMAGp956C1376	AA398981		KIAA0874	"KIAA0874 protein"
581	68811	IMAGp956A1176	AA398975		OFD1	"oral-facial-digital syndrom
582	69171	IMAGp956P1176	AA417226		ACE	"angiotensin I converting en
583	69053	IMAGp956K1376	AA399077		KIAA1324	"KIAA1324 protein"
584	68982	IMAGp956H1476	AA417319		CDC14B	"CDC14 cell division cycle 1
585	70050	IMAGp956E0279	AA625899		FLJ14351	"hypothetical protei
586	70051	IMAGp956E0379	AA621310		FOXE3	"forkhead box E3" RZPD
587	70222	IMAGp956L0679	AA421986		INPP5D	"inositol polyphosphate-5-ph
588	69984	IMAGp956B0879	AA419362		CETN3	"centrin, EF-hand protein, 3
589	70004	IMAGp956C0479	AA634287		PRKAR2A	"protein kinase, cAMP-depend
590	70078	IMAGp956F0679	AA419345		METAP1	"methionyl aminopeptidase 1"
591	70129	IMAGp956H0979	AA625737		GCAT	"glycine C-acetyltransferase
592	70249	IMAGp956M0979	AA626368		CTNND1	"catenin (cadherin-associate
593	71028	IMAGp956M2081	AA437224		NKX3-1	"NK3 transcription factor re
594	70765	IMAGp956B2181	AA436720		NCBP1	"nuclear cap binding protein
595	70960	IMAGp956J2481	AA442703		DKFZP564D0462	"hypothetical protei
596	71201	IMAGp956E0182	AA418718		MAGEL2	"MAGE-like 2" RZPD Hs.1
597	70815	IMAGp956D2381	AA442221		LOC51003	"CGI-125 protein"
598	70768	IMAGp956B2481	AA436902		DJ1220B.2	"hypothetical protei
599	71417	IMAGp956N0182	AA425089		CLOCK	"clock homolog (mouse)" RZPD

600	71154	IMAGp956C0282	AA418901	EIF3S9	"eukaryotic translation init
601	72049	IMAGp956H0984	AA431779	EPPB9	"B9 protein" RZPD Hs.1
602	71954	IMAGp956D1084	AA431942	SIAT6	"sialyltransferase 6 (N-acet
603	71910	IMAGp956B1484	AA431912	SPRY2	"sprouty homolog 2 (Drosophi
604	72055	IMAGp956H1584	AA446604	AD-003	"AD-003 protein" RZPD
605	72123	IMAGp956K1184	AA431245	ANKRD7	"ankyrin repeat domain 7"
606	72029	IMAGp956G1384	AA429586	DKFZp762P2111	"hypothetical protei
607	72151	IMAGp956L1584	AA447588	DBY	"DEAD/H (Asp-Glu-Ala-Asp/His
608	72056	IMAGp956H1684	AA443854	TACSTD1	"tumor-associated calcium si
609	72874	IMAGp956J1886	AA460850	COX7A1	"cytochrome c oxidase subuni
610	72994	IMAGp956O1886	AA454207	ABHD2	"abhydrolase domain containi
611	72708	IMAGp956C2086	AA459673	SMC4L1	"SMC4 structural maintenance
612	72900	IMAGp956K2086	AA460285	FKBP6	"FK506 binding protein 6, 36
613	72683	IMAGp956B1986	AA460934	BC-2	"putative breast adenocarcin
614	72755	IMAGp956E1986	AA459707	SLC2A5	"solute carrier family 2 (fa
615	72948	IMAGp956M2086	AA460333	KIAA1466	"KIAA1466 protein"
616	72853	IMAGp956I2186	AA460595	DKFZP434B204	"DKFZP434B204 protei
617	73552	IMAGp956F2488	AA490306	FPGT	"fucose-1-phosphate guanylyl
618	74153	IMAGp956P0189	AA496794	PUS1	"pseudouridylate synthase 1"
619	73917	IMAGp956F0589	AA488863	FLJ11125	"hypothetical protei
620	74110	IMAGp956N0689	AA488177	DCTN4	"dynactin 4 (p62)" RZPD
621	74106	IMAGp956N0289	AA496839	EIF3S4	"eukaryotic translation init
622	74108	IMAGp956N0489	AA496840	RPL7A	"ribosomal protein L7a" RZPD
623	73992	IMAGp956I0889	AA504146	ELAC2	"elaC homolog 2 (E. coli)"
624	73853	IMAGp956C1389	AA488801	CLK3	"CDC-like kinase 3" RZPD
625	77042	IMAGp956H1097	AA889705	OXCT2	"3-oxoacid CoA transferase 2
626	76875	IMAGp956A1197	AA868278	TPX1	"testis specific protein 1 (
627	76998	IMAGp956F1497	AA844930	GP2	"glycoprotein 2 (zymogen gra
628	76903	IMAGp956B1597	AA835004	MOT8	"hypothetical protein MOT8"
629	77235	IMAGp956P1197	AA843704	C21orf8	"chromosome 21 open reading
630	76902	IMAGp956B1497	AA844831	CPA2	"carboxypeptidase A2 (pancre
631	76976	IMAGp956E1697	AA813015	KLRD1	"killer cell lectin-like rec
632	76954	IMAGp956D1897	AA844864	REG1B	"regenerating islet-derived
633	81205	IMAGp956E21108	AI016769	DKFZP434P1735	"hypothetical protei
634	81447	IMAGp956O23108	AA999850	NX17	"kidney-specific membrane pr
635	81144	IMAGp956C08108	AI000434	GLTP	"glycolipid transfer protein
636	81409	IMAGp956N09108	AI015552	SGK2	"serum/glucocorticoid regula
637	81428	IMAGp956O04108	AI075049	AQP2	"aquaporin 2 (collecting duc
638	81238	IMAGp956G06108	AI081046	BYSL	"bystin-like" RZPD Hs.1
639	81746	IMAGp956L10109	AI018069	IRAK3	"interleukin-1 receptor-asso
640	81727	IMAGp956K15109	AI016765		"ESTs, Moderately similar to
641	84979	IMAGp956C03118	AI126211	FLJ11342	"hypothetical protei
642	85199	IMAGp956L07118	AI125323	RSHL1	"radial spokehead-like 1"
643	85493	IMAGp956H13119	AI193306	ICAM4	"intercellular adhesion mole
644	85543	IMAGp956J15119	AI193053	LOC56965	"hypothetical protei
645	85056	IMAGp956F08118	AI143924	FLJ10786	"hypothetical protei
646	85514	IMAGp956I10119	AI186169	FLJ12526	"hypothetical protei
647	85569	IMAGp956K17119	AI143954	C1orf34	"chromosome 1 open reading f
648	85666	IMAGp956O18119	AI191074	UPF3B	"similar to yeast Upf3, vari
649	89889	IMAGp956O17130	AI333565	DGKH	"diacylglycerol kinase, eta"
650	89746	IMAGp956I18130	AI334188	LOC51333	"mesenchymal stem ce
651	89703	IMAGp956G23130	AI333429	KIAA1615	"KIAA1615 protein"
652	89830	IMAGp956M06130	AI339609	KCNE2	"potassium voltage-gated cha
653	89891	IMAGp956O19130	AI351311		"ESTs, Highly similar to dJ1
654	89677	IMAGp956F21130	AI332905	FLJ10498	"hypothetical protei
655	89735	IMAGp956I07130	AI350748	PCDH19	"protocadherin 19" RZPD
656	89592	IMAGp956C08130	AI333844	KCNIP1	"Kv channel interacting prot
657	94775	IMAGp956K07143	AI479375	ITGA2B	"integrin, alpha 2b (platele
658	95138	IMAGp956J10144	AI633839	FLJ21135	"hypothetical protei
659	95237	IMAGp956N13144	AI918804	WDR4	"WD repeat domain 4" RZPD

660	95120	IMAGp956I16144	AI565972		GRIN2D	"glutamate receptor, ionotro
661	95116	IMAGp956I12144	AI582219		MGC10796	"hypothetical protei
662	95093	IMAGp956H13144	AI916727		NEUROD2	"neurogenic differentiation
663	94954	IMAGp956B18144	AI695885		HDC	"histidine decarboxylase"
664	95173	IMAGp956K21144	AI564953		HAP1	"huntingtin-associated prote
665	100303	IMAGp956A15158	AI910449		DKFZP564O0823	"DKFZP564O0823 prote
666	100521	IMAGp956J17158	AI768802		WWP2	"Nedd-4-like ubiquitin-prote
667	100628	IMAGp956O04158	AI761253		C21orf52	"chromosome 21 open
668	100318	IMAGp956B06158	AI831744		FZD9	"frizzled homolog 9 (Drosoph
669	100355	IMAGp956C19158	AI765117		FLJ10647	"hypothetical protei
670	100480	IMAGp956H24158	AI858088		ALOX15B	"arachidonate 15-lipoxygenas
671	100295	IMAGp956A07158	AI738634		TNFRSF4	"tumor necrosis factor recep
672	100655	IMAGp956P07158	AI819741		BCAS4	"breast carcinoma amplified
673	103278	IMAGp956M14165	AI357407		GUCY2C	"guanylate cyclase 2C (heat
674	103209	IMAGp956J17165	AI284076		RPF1	"RNA processing factor 1"
675	103293	IMAGp956N05165	AI270709		DCLRE1B	"DNA cross-link repair 1B (P
676	103151	IMAGp956H07165	AI282890		EFNA3	"ephrin-A3" RZPD Hs.3
677	103259	IMAGp956L19165	AI280589		FLJ10565	"hypothetical protei
678	103216	IMAGp956J24165	AI280656		LZK1	"C3HC4-type zinc finger prot
679	103032	IMAGp956C08165	AI281892		MUC2	"mucin 2, intestinal/trachea
680	103344	IMAGp956P08165	AI355885		KIAA1005	"KIAA1005 protein"
681	106227	IMAGp956H11173	AI610252		FLJ12765	"hypothetical protei
682	106349	IMAGp956M13173	AI801727		GALNT11	"GALNAC-T11" RZPD Hs.9
683	106098	IMAGp956C02173	AI919501		MTAP	"methylthioadenosine phospho
684	106356	IMAGp956M20173	AI919582		MJD	"Machado-Joseph disease (spi
685	106302	IMAGp956K14173	AI537934		KIAA1271	"KIAA1271 protein"
686	106256	IMAGp956I16173	AI610888		KIAA1679	"KIAA1679 protein"
687	106191	IMAGp956F23173	AI825946		CTXL	"cortical thymocyte receptor
688	106339	IMAGp956M03173	AI570373		DDX27	"DEAD/H (Asp-Glu-Ala-Asp/His
689	110642	IMAGp956P10184	AI862711		RGS13	"regulator of G-protein sign
690	110452	IMAGp956H12184	AI694095		KIAA1372	"KIAA1372 protein"
691	110442	IMAGp956H02184	AI696847		C10orf3	"chromosome 10 open reading
692	110341	IMAGp956C21184	AI866617		FLJ13204	"hypothetical protei
693	110645	IMAGp956P13184	AI697014		FLJ13593	"hypothetical protei
694	110435	IMAGp956G19184	AI863388		KIAA1446	"KIAA1446 protein"
695	110581	IMAGp956M21184	AI635302		KIAA1061	"KIAA1061 protein"
696	110392	IMAGp956E24184	AI635899		GG2-1	"TNF-induced protein" RZPD
697	5197	956077		""	GS	Hs.388 Unknown Gene
698	5186	1227385	EPHB3	"EphB3"	GS	Hs.2913 Known Gene
699	5141	375220	MEOX2	"mesenchyme homeo box 2 (growth arrest-speci		
700	5178	428236	SFRP1	"secreted frizzled-related protein 1" GS		
701	5187	955697	FRG1	"FSHD region gene 1" GS Hs.203772		
702	5176	1996511	TPM4	"tropomyosin 4" GS Hs.250641 Know		
703	5167	1998594	FOSB	"FBJ murine osteosarcoma viral oncogene homo		
704	5132	1997792	TIP-1	"Tax interaction protein 1" GS Hs.1		
705	5714	374875		"ESTs" GS Hs.171171 Unknown Gene		
706	5750	1432207		"ESTs" GS Hs.315562 Unknown Gene		
707	5775	663057		"ESTs" GS Hs.21658 Unknown Gene		
708	5740	623743	ZNF189	"zinc finger protein 189" GS Hs.5		
709	5774	795187	KCNE3	"potassium voltage-gated channel, Isk-relate		
710	5715	2054426		"ESTs" GS Hs.44099 Unknown Gene		
711	5764	2054436		"ESTs" GS Hs.269689 Unknown Gene		
712	5729	1229825		"" GS Hs.7985 Unknown Gene		
713	6039	61183		"ESTs" GS Hs.35828 Unknown Gene		
714	5992	2247386		"ESTs" GS Hs.61119 Unknown Gene		
715	5981	3253716		"" GS Hs.341567 Unknown Gene		
716	6005	435011	CEACAM6	"carcinoembryonic antigen-related cell adhes		
717	6028	532490		"ESTs" GS Hs.116417 Unknown Gene		
718	6052	1794545		"" GS Hs.306591 Unknown Gene		
719	6041	1760716	SSBP4	"single stranded DNA binding protein 4" GS		

720	6018	2253878	DKFZp434I1610		"hypothetical protein DKFZp434I1610"			
721	67845	IMAGp956I0573	AA280742		THTPA	"thiamine triphosphatase"		
722	67870	IMAGp956J0673	AA235706		TAF10	"TAF10 RNA polymerase II, TA		
723	67800	IMAGp956G0873	AA281621		BIRC2	"baculoviral IAP repeat-cont		
724	67825	IMAGp956H0973	AA280304		EMD	"emerin (Emery-Dreifuss musc		
725	67823	IMAGp956H0773	AA280303		FMR1	"fragile X mental retardatio		
726	67656	IMAGp956A0873	AA281007		FLJ11198	"hypothetical protei		
727	67874	IMAGp956J1073	AA284522		PIAS3	"protein inhibitor of activa		
728	67922	IMAGp956L1073	AA235733		ZNF32	"zinc finger protein 32 (KOX		
729	6757	2254021	FLJ31818		"hypothetical protein FLJ31818"	GS		
730	6793	2245848		""	GS	Unknown Gene		
731	6759	2009069		"ESTs"	GS	Hs.98197	Unknown Gene	
732	6795	1923289	GJB5	"gap junction protein, beta 5 (connexin 31.1				
733	6758	2044495		""	GS	Hs.12755	Unknown Gene	
734	6794	2292949	KIAA1805		"KIAA1805 protein"	GS	Hs.2	
735	6808	394219	FLJ25348		"hypothetical protein FLJ25348"	GS		
736	6761	4003220	LOC115509		"hypothetical protein BC014000"	GS		
737	40345	IMAGp956O091	T67269		""	RZPD	Unknown Gene	
738	40183	IMAGp956H151	T66397	COG8	"component of oligomeric golgi compl			
739	40657	IMAGp956L092	R37600	KIAA1889	"KIAA1889 protein"	RZPD		
740	40526	IMAGp956F222	T69758		"Homo sapiens cDNA FLJ32216 fis, clo			
741	40118	IMAGp956E221	R15709		""	RZPD	Unknown Gene	
742	40385	IMAGp956A012	T77103	LOC58504	"hypothetical protein from c			
743	40622	IMAGp956J222	T77818	DKFZp434I1930	"hypothetical protein DKFZp4			
744	40991	IMAGp956J073	T81332	SURF4	"surfeit 4"	RZPD	Hs.284296	
745	49616	IMAGp956A1626	H27374	KIAA1798	"KIAA1798 protein"	RZPD		
746	49740	IMAGp956F2026	H06508	MGC20727	"hypothetical protein MGC207			
747	50194	IMAGp956I1827	H09520	TTYH2	"tweety homolog 2 (Drosophila)"	RZPD		
748	50683	IMAGp956N0328	H15416	MEGF10	"MEGF10 protein"	RZPD	Hs.2	
749	50346	IMAGp956P0227	H11343	FLJ14594	"hypothetical protein FLJ145			
750	50038	IMAGp956C0627	H09166	CECR6	"cat eye syndrome chromosome region,			
751	50636	IMAGp956L0428	H17649	LOC115294	"similar to hypothetical pro			
752	50594	IMAGp956J1028	H16803	MGC3251	"hypothetical protein MGC3251"	RZPD		
753	62431	IMAGp956G1559	W51748	FLJ14437	"myopalladin"	RZPD	Hs.5	
754	62483	IMAGp956I1959	W63783	TM6SF2	"transmembrane 6 superfamily member			
755	62921	IMAGp956L0160	W72893	HCCA2	"HCCA2 protein"	RZPD	Hs.19223	
756	62802	IMAGp956G0260	W72897	MGC31963	"hypothetical protein MGC319			
757	62580	IMAGp956M2059	W57724		"ESTs, Weakly similar to hypothetica			
758	62629	IMAGp956O2159	W02121	FLJ12525	"hypothetical protein FLJ125			
759	62928	IMAGp956L0860	W71997	IMAGE3451454	"GRASP protein"	RZPD	Hs.3	
760	62718	IMAGp956C1460	W73386	CIDEA	"cell death-inducing DFFA-like effec			
761				""	Unknown Gene			
762	73851	IMAGp956C1189	AA504137		LAGY	"homeodomain only protein"		
763	74659	IMAGp956E0391	AA563651		MGC13038	"hypothetical protei		
764	74742	IMAGp956H1491	AA505930		DKFZP434A1315	"hypothetical protei		
765	74167	IMAGp956P1589	AA521500		C21orf67	"chromosome 21 open		
766	74001	IMAGp956I1789	AA504464		PCCX2	"protein containing CXXC dom		
767	75401	IMAGp956D0193	AA608857		SSTK	"serine/threonine protein ki		
768	75564	IMAGp956J2093	AA644487			"ESTs, Moderately similar to		
769	98240	IMAGp956K16152	AI694562		COL4A3	"collagen, type IV, alpha 3		
770	98443	IMAGp956D03153	AI796896		HOXA3	"homeo box A3"	RZPD	Hs.2
771	100667	IMAGp956P19158	AI765426		NUDT12	"nudix (nucleoside diphospha		
772	101014	IMAGp956O06159	AI935449		TIMM8A	"translocase of inner mitoch		
773	99621	IMAGp956E05156	AI740592		KCNK17	"potassium channel, subfamil		
774	99878	IMAGp956O22156	AI740827		CDH12	"cadherin 12, type 2 (N-cadh		
775	101420	IMAGp956P04160	H23529	FLJ14779	"hypothetical protein FLJ147			
776	101352	IMAGp956M08160	R12994	AF311304	"hypothetical protein AF3113			
777	41021	IMAGp956K133	T77824	TMTSP	"TMTSP for transmembrane molecule wi			
778	41023	IMAGp956K153	T78903	LOC93556	"hypothetical protein BC0112			
779	40813	IMAGp956B213	T71279		"Homo sapiens, clone IMAGE:3464710,			

EXHIBIT K

```
MEGABLAST 1.2.3-Paracel [2001-11-20]
Reference:
Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000),
"A greedy algorithm for aligning DNA sequences",
J Comput Biol 2000; 7(1-2):203-14.
Database: Homo_sapiens.latestgp.masked.fa
           33,840 sequences; 200,810,911,373 total letters

Query= LEX0317seqid1
         (3423 letters)

                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

AC008682.6.1.217221                                                 492  e-136
AC010424.9.1.192282                                                 476  e-131
AC084373.24.1.184876                                                 92  3e-15

>AC008682.6.1.217221
          Length = 217221

 Score =  492 bits (248), Expect = e-136
 Identities = 248/248 (100%)
 Strand = Plus / Minus


Query: 412    gcctgcgatggtgatcactggggtccccactgcaccagccggtgccagtgcaaaaatggg 471
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 105219 gcctgcgatggtgatcactggggtccccactgcaccagccggtgccagtgcaaaaatggg 105160


Query: 472    gctctgtgcaaccccatcaccggggcttgccactgtgctgcgggcttccggggctggcgc 531
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 105159 gctctgtgcaaccccatcaccggggcttgccactgtgctgcgggcttccggggctggcgc 105100


Query: 532    tgcgaggaccgctgtgagcagggcacctatggtaacgactgtcatcagagatgccagtgc 591
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 105099 tgcgaggaccgctgtgagcagggcacctatggtaacgactgtcatcagagatgccagtgc 105040


Query: 592    cagaatggagccacctgcgaccacgtcacgggggaatgccgctgcccaccaggatacacc 651
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 105039 cagaatggagccacctgcgaccacgtcacgggggaatgccgctgcccaccaggatacacc 104980


Query: 652    ggagcctt 659
              ||||||||
Sbjct: 104979 ggagcctt 104972


 Score =  476 bits (240), Expect = e-131
 Identities = 240/240 (100%)
 Strand = Plus / Minus


Query: 2490  agctggtgttatcatagttggaaatctgaacagcttaagccgaaccagtactgctctccc 2549
```

```
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56296 agctggtgttatcatagttggaaatctgaacagcttaagccgaaccagtactgctctccc 56237


Query: 2550  tgctgattcctaccagatcggggccattgcaggcatcatcattcttgtcctagttgttct 2609
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56236 tgctgattcctaccagatcggggccattgcaggcatcatcattcttgtcctagttgttct 56177


Query: 2610  cttcctactggcattgttcattatttatagacacaagcagaagggaaaggaatcaagcat 2669
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56176 cttcctactggcattgttcattatttatagacacaagcagaagggaaaggaatcaagcat 56117


Query: 2670  gccagcagttacctacacccctgctatgagggtcgtcaatgcagattataccatttcagg 2729
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 56116 gccagcagttacctacacccctgctatgagggtcgtcaatgcagattataccatttcagg 56057


 Score =  416 bits (210), Expect = e-113
 Identities = 213/214 (99%)
 Strand = Plus / Minus


Query: 917   ggtgccaggatgagtgtcctgttgggacctatggcgttctctgtgctgagacctgccagt 976
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 91362 ggtgccaggatgagtgtcctgttgggacctatggcgttctctgtgctgagacctgccagt 91303


Query: 977   gtgtcaacggagggaagtgttaccacgtgagcggcgcatgcctctgtgaagcaggctttg 1036
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 91302 gtgtcaacggagggaagtgttaccacgtgagcggcgcatgcctctgtgaagcaggctttg 91243


Query: 1037  ctggcgagcgctgcgaagcacgcctgtgtcctgaggggctctacggcatcaaatgtgaca 1096
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 91242 ctggcgagcgctgcgaagcacgcctgtgtcctgaggggctctacggcatcaaatgtgaca 91183


Query: 1097  aacggtgtccctgccacttggaaaacactcatag 1130
              |||||||||||||||| ||||||||||||||
Sbjct: 91182 aacggtgtccctgccacctggaaaacactcatag 91149


 Score =  414 bits (209), Expect = e-112
 Identities = 209/209 (100%)
 Strand = Plus / Minus


Query: 3024  agacctgggaaagaattctgaatataattcaagtaactgctccctaagcagttctgagaa 3083
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46352 agacctgggaaagaattctgaatataattcaagtaactgctccctaagcagttctgagaa 46293


Query: 3084  cccatatgccactattaaagacccacctgtacttatcccgaaaagctcagagtgtggtta 3143
```

```
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46292 cccatatgccactattaaagacccacctgtacttatcccgaaaagctcagagtgtggtta 46233


Query: 3144  tgtggagatgaaatcgccggcacgaagagattccccatatgcagagatcaataactcaac 3203
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46232 tgtggagatgaaatcgccggcacgaagagattccccatatgcagagatcaataactcaac 46173


Query: 3204  ttcagccaacaggaatgtctatgaagttg 3232
            |||||||||||||||||||||||||||||
Sbjct: 46172 ttcagccaacaggaatgtctatgaagttg 46144


 Score =  353 bits (178), Expect = 5e-94
 Identities = 178/178 (100%)
 Strand = Plus / Minus


Query: 1129  agctgtcaccccatgtctggagagtgtgcctgcaagccgggctggtcaggactctactgt 1188
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 84114 agctgtcaccccatgtctggagagtgtgcctgcaagccgggctggtcaggactctactgt 84055


Query: 1189  aatgagacatgttctcctggattctacggggaagcttgccagcagatctgcagctgccaa 1248
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 84054 aatgagacatgttctcctggattctacggggaagcttgccagcagatctgcagctgccaa 83995


Query: 1249  aatggggcagactgtgacagtgtgactggaaagtgcacctgtgccccaggattcaaag 1306
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 83994 aatggggcagactgtgacagtgtgactggaaagtgcacctgtgccccaggattcaaag 83937


 Score =  333 bits (168), Expect = 4e-88
 Identities = 168/168 (100%)
 Strand = Plus / Minus


Query: 1424  caggctggcacggggtggactgctccatcagatgtcccagtggcacatggggctttggct 1483
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 81709 caggctggcacggggtggactgctccatcagatgtcccagtggcacatggggctttggct 81650


Query: 1484  gtaacttaacatgccagtgcctcaacgggggagcctgcaacaccctggacgggacctgca 1543
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 81649 gtaacttaacatgccagtgcctcaacgggggagcctgcaacaccctggacgggacctgca 81590


Query: 1544  cgtgtgcacctggatggcgcggggagaaatgcgaacttccctgccagg 1591
            ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 81589 cgtgtgcacctggatggcgcggggagaaatgcgaacttccctgccagg 81542
```

Score = 299 bits (151), Expect = 6e-78
Identities = 151/151 (100%)
Strand = Plus / Minus

```
Query: 1690  tcaggtgtccactgtgacagcgtgtgtgctgagggacgctggggccccaactgctccctg 1749
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 68391 tcaggtgtccactgtgacagcgtgtgtgctgagggacgctggggccccaactgctccctg 68332


Query: 1750  ccctgctactgtaaaaatggggcttcatgctcccctgatgatggcatctgcgagtgtgca 1809
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 68331 ccctgctactgtaaaaatggggcttcatgctcccctgatgatggcatctgcgagtgtgca 68272


Query: 1810  ccaggcttccgaggcaccacttgtcagagga 1840
              |||||||||||||||||||||||||||||
Sbjct: 68271 ccaggcttccgaggcaccacttgtcagagga 68241
```

Score = 278 bits (140), Expect = 2e-71
Identities = 140/140 (100%)
Strand = Plus / Minus

```
Query: 780   gggcacagtgtgtggtcagccttgccccgagggtcgctttggaaagaactgttcccaaga 839
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 99205 gggcacagtgtgtggtcagccttgccccgagggtcgctttggaaagaactgttcccaaga 99146


Query: 840   atgccagtgccataatggagggacgtgtgatgctgccacaggccaatgtcattgcagtcc 899
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 99145 atgccagtgccataatggagggacgtgtgatgctgccacaggccaatgtcattgcagtcc 99086


Query: 900   aggatacacaggggaacggt 919
              ||||||||||||||||||||
Sbjct: 99085 aggatacacaggggaacggt 99066
```

Score = 268 bits (135), Expect = 2e-68
Identities = 135/135 (100%)
Strand = Plus / Minus

```
Query: 1841  tctgctcccctggtttttatgggcatcgctgcagccagacatgcccacagtgcgttcaca 1900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 67063 tctgctcccctggtttttatgggcatcgctgcagccagacatgcccacagtgcgttcaca 67004


Query: 1901  gcagcgggccctgccaccacatcaccggcctgtgtgactgcttgcctggcttcacaggcg 1960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 67003 gcagcgggccctgccaccacatcaccggcctgtgtgactgcttgcctggcttcacaggcg 66944
```

```
Query: 1961  ccctctgcaatgaag 1975
             ||||||||||||||
Sbjct: 66943 ccctctgcaatgaag 66929
```

Score = 260 bits (131), Expect = 5e-66
Identities = 131/131 (100%)
Strand = Plus / Minus

```
Query: 1974  agtgtgtcccagtggcagatttgggaaaaactgtgcaggaatttgtacctgcaccaacaa 2033
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 66391 agtgtgtcccagtggcagatttgggaaaaactgtgcaggaatttgtacctgcaccaacaa 66332

Query: 2034  cggaacctgtaaccccattgacagatcttgtcagtgttaccccggttggattggcagtga 2093
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 66331 cggaacctgtaaccccattgacagatcttgtcagtgttaccccggttggattggcagtga 66272

Query: 2094  ctgctctcaac 2104
             ||||||||||||
Sbjct: 66271 ctgctctcaac 66261
```

Score = 260 bits (131), Expect = 5e-66
Identities = 131/131 (100%)
Strand = Plus / Minus

```
Query: 2726  caggaacccttcctcacagcaatggtggaaacgctaatagccactacttcaccaatccca 2785
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54197 caggaacccttcctcacagcaatggtggaaacgctaatagccactacttcaccaatccca 54138

Query: 2786  gttaccacacgctcacccagtgtgccacatcccctcacgtcaacaacagggacaggatga 2845
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 54137 gttaccacacgctcacccagtgtgccacatcccctcacgtcaacaacagggacaggatga 54078

Query: 2846  ctgtcacgaag 2856
             ||||||||||||
Sbjct: 54077 ctgtcacgaag 54067
```

Score = 258 bits (130), Expect = 2e-65
Identities = 130/130 (100%)
Strand = Plus / Minus

```
Query: 2105  catgtccacctgcccactggggcccaaactgcatccacacgtgcaactgccataatggag 2164
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 63311 catgtccacctgcccactggggcccaaactgcatccacacgtgcaactgccataatggag 63252
```

```
Query: 2165  ctttctgcagcgcctacgatggggaatgtaaatgcactcctggctggacagggctctact 2224
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 63251 ctttctgcagcgcctacgatggggaatgtaaatgcactcctggctggacagggctctact 63192


Query: 2225  gcactcagag 2234
             ||||||||||
Sbjct: 63191 gcactcagag 63182
```

Score = 256 bits (129), Expect = 8e-65
Identities = 129/129 (100%)
Strand = Plus / Minus

```
Query: 2363  agtgcccttcaggaacatatggctatggctgtcgccagatatgtgattgtctgaacaact 2422
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 58752 agtgcccttcaggaacatatggctatggctgtcgccagatatgtgattgtctgaacaact 58693


Query: 2423  ccacctgcgaccacatcactgggacctgttactgcagccccggatggaagggagcgagat 2482
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 58692 ccacctgcgaccacatcactgggacctgttactgcagccccggatggaagggagcgagat 58633


Query: 2483  gtgatcaag 2491
             |||||||||
Sbjct: 58632 gtgatcaag 58624
```

Score = 256 bits (129), Expect = 8e-65
Identities = 129/129 (100%)
Strand = Plus / Minus

```
Query: 2855  agtcaaaaaacaatcaactgtttgtgaatcttaaaaatgtgaaccctgggaagagaggcc 2914
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 52654 agtcaaaaaacaatcaactgtttgtgaatcttaaaaatgtgaaccctgggaagagaggcc 52595


Query: 2915  ctgtgggggactgcactgggacattgccggctgactggaaacatggcggctacctcaacg 2974
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 52594 ctgtgggggactgcactgggacattgccggctgactggaaacatggcggctacctcaacg 52535


Query: 2975  agctcggtg 2983
             |||||||||
Sbjct: 52534 agctcggtg 52526
```

Score = 256 bits (129), Expect = 8e-65
Identities = 129/129 (100%)
Strand = Plus / Minus

```
Query: 2234  gatgtcctctagggttttatggaaaagattgtgcactgatatgccaatgtcaaaacggag 2293
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61011 gatgtcctctagggttttatggaaaagattgtgcactgatatgccaatgtcaaaacggag 60952


Query: 2294  ctgactgcgaccacatttctgggcagtgtacttgccgcactggattcatgggacggcact 2353
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 60951 ctgactgcgaccacatttctgggcagtgtacttgccgcactggattcatgggacggcact 60892


Query: 2354  gtgagcaga 2362
              |||||||||
Sbjct: 60891 gtgagcaga 60883
```

17

Score = 242 bits (122), Expect = 1e-60
Identities = 122/122 (100%)
Strand = Plus / Minus

```
Query: 1306  ggaattgactgctctaccccatgccctctgggaacctatgggataaactgttcctctcgc 1365
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 82630 ggaattgactgctctaccccatgccctctgggaacctatgggataaactgttcctctcgc 82571


Query: 1366  tgtggctgtaaaaatgatgcagtctgctctcctgtggacgggtcttgtacttgcaaggca 1425
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 82570 tgtggctgtaaaaatgatgcagtctgctctcctgtggacgggtcttgtacttgcaaggca 82511


Query: 1426  gg 1427
              ||
Sbjct: 82510 gg 82509
```

10

Score = 242 bits (122), Expect = 1e-60
Identities = 122/122 (100%)
Strand = Plus / Minus

```
Query: 660    ctgtgaggatctttgtcctcctggtaaacatggtccacagtgtgagcagagatgcccttg 719
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 103074 ctgtgaggatctttgtcctcctggtaaacatggtccacagtgtgagcagagatgcccttg 103015


Query: 720    tcaaaatggaggagtgtgtcatcacgtcactggagaatgctcttgcccttctggctggat 779
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 103014 tcaaaatggaggagtgtgtcatcacgtcactggagaatgctcttgcccttctggctggat 102955


Query: 780    gg 781
               ||
Sbjct: 102954 gg 102953
```

6

```
 Score =  230 bits (116), Expect = 5e-57
 Identities = 116/116 (100%)
 Strand = Plus / Minus


Query: 1      atggttatttctttgaactcatgcctgagctttatttgtttattgttatgccactggatt 60
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 170441 atggttatttctttgaactcatgcctgagctttatttgtttattgttatgccactggatt 170382


Query: 61     gggacagcatcacctctgaatcttgaagaccctaatgtgtgtagccactgggaaag 116
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 170381 gggacagcatcacctctgaatcttgaagaccctaatgtgtgtagccactgggaaag 170326


 Score =  218 bits (110), Expect = 2e-53
 Identities = 165/192 (85%)
 Strand = Plus / Minus


Query: 3232  gaacctacagtgagtgttgtccaaggagtattcagcaataatgggcgtctctcccaggat 3291
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44624 gaacctacagtgagtgttgtccaaggagtattcagcaataatgggcgtctctcccaggat 44565


Query: 3292  ccatatgacctcccaaagaacagtcacatcccttgtcattatgacctgctgccagtccga 3351
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44564 ccatatgacctcccaaagaacagtcacatcccttgtcattatgacctgctgccagtccga 44505


Query: 3352  gacagttcatcctcccctaagcaagaggacagtggaggtagcagcagcaacagcagcagc 3411
              |||||||||||||||||||||||||||||||||| |||
Sbjct: 44504 gacagttcatcctcccctaagcaagaggacagtggtggtnnnnnnnnnnnnnnnnnnnn 44445


Query: 3412  agcagtgaatga 3423
                   |||||||
Sbjct: 44444 nnnnntgaatga 44433


 Score =  208 bits (105), Expect = 2e-50
 Identities = 108/109 (99%)
 Strand = Plus / Minus


Query: 1588  caggatggcacgtacgggctgaactgtgctgagcgctgcgactgcagccacgcagatggc 1647
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 79083 caggatggcacgtacgggctgaactgtgctgagcgctgcgactgcagccacgcagatggc 79024


Query: 1648  tgccaccctaccacgggccattgccgctgcctcccgggatggtcaggtg 1696
              |||||||||||||||||||||||||||||||| ||||||||||||
```

```
Sbjct: 79023 tgccaccctaccacgggccattgccgctgcctccccggatggtcaggtg 78975


 Score =  206 bits (104), Expect = 7e-50
 Identities = 104/104 (100%)
 Strand = Plus / Minus

Query: 115    agctactcagtgactgtgcaagagtcatacccacatccctttgatcaaatttactacacg 174
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 162632 agctactcagtgactgtgcaagagtcatacccacatccctttgatcaaatttactacacg 162573


Query: 175    agctgcactgacattctaaactggtttaaatgcacgcggcacag 218
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct: 162572 agctgcactgacattctaaactggtttaaatgcacgcggcacag 162529


 Score =  204 bits (103), Expect = 3e-49
 Identities = 103/103 (100%)
 Strand = Plus / Minus


Query: 217    agagtcagctatcggacagcctatcgacatggggagaagactatgtataggcgcaagtct 276
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 161222 agagtcagctatcggacagcctatcgacatggggagaagactatgtataggcgcaagtct 161163


Query: 277    cagtgttgtcctggattttatgaaagcggggaaatgtgtgtcc 319
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct: 161162 cagtgttgtcctggattttatgaaagcggggaaatgtgtgtcc 161120


 Score =  184 bits (93), Expect = 2e-43
 Identities = 93/93 (100%)
 Strand = Plus / Minus


Query: 320    cccactgtgctgataaatgtgtccatggtcgctgtattgctccaaacacctgtcagtgtg 379
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 131840 cccactgtgctgataaatgtgtccatggtcgctgtattgctccaaacacctgtcagtgtg 131781


Query: 380    agcctggctggggagggaccaactgctccagtg 412
              |||||||||||||||||||||||||||||||||
Sbjct: 131780 agcctggctggggagggaccaactgctccagtg 131748


 Score = 91.7 bits (46), Expect = 3e-15
 Identities = 46/46 (100%)
 Strand = Plus / Minus
```

```
Query: 2980  ggtgcttttggacttgacagaagctatatgggaaaatccttaaaag 3025
              |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 47186 ggtgcttttggacttgacagaagctatatgggaaaatccttaaaag 47141
```

>AC010424.9.1.192282
Length = 192282

Score = 476 bits (240), Expect = e-131
Identities = 240/240 (100%)
Strand = Plus / Plus

```
Query: 2490  agctggtgttatcatagttggaaatctgaacagcttaagccgaaccagtactgctctccc 2549
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 14373 agctggtgttatcatagttggaaatctgaacagcttaagccgaaccagtactgctctccc 14432

Query: 2550  tgctgattcctaccagatcggggccattgcaggcatcatcattcttgtcctagttgttct 2609
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 14433 tgctgattcctaccagatcggggccattgcaggcatcatcattcttgtcctagttgttct 14492

Query: 2610  cttcctactggcattgttcattatttatagacacaagcagaagggaaaggaatcaagcat 2669
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 14493 cttcctactggcattgttcattatttatagacacaagcagaagggaaaggaatcaagcat 14552

Query: 2670  gccagcagttacctacacccctgctatgagggtcgtcaatgcagattataccatttcagg 2729
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 14553 gccagcagttacctacacccctgctatgagggtcgtcaatgcagattataccatttcagg 14612
```

Score = 414 bits (209), Expect = e-112
Identities = 209/209 (100%)
Strand = Plus / Plus

```
Query: 3024  agacctgggaaagaattctgaatataattcaagtaactgctccctaagcagttctgagaa 3083
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24317 agacctgggaaagaattctgaatataattcaagtaactgctccctaagcagttctgagaa 24376

Query: 3084  cccatatgccactattaaagacccacctgtacttatcccgaaaagctcagagtgtggtta 3143
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24377 cccatatgccactattaaagacccacctgtacttatcccgaaaagctcagagtgtggtta 24436

Query: 3144  tgtggagatgaaatcgccggcacgaagagattccccatatgcagagatcaataactcaac 3203
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 24437 tgtggagatgaaatcgccggcacgaagagattccccatatgcagagatcaataactcaac 24496

Query: 3204  ttcagccaacaggaatgtctatgaagttg 3232
              |||||||||||||||||||||||||||||
```